Data Reusers' Trust Development

Ayoung Yoon

School of Informatics and Computing, Department of Library and Information Science, Indiana University Purdue University Indianapolis, 535 W. Michigan Street, IT 563, Indianapolis, IN 46202-3103.
E-mail: ayyoon@iupui.edu

Data reuse refers to the secondary use of data—not for its original purpose but for studying new problems. Although reusing data might not yet be the norm in every discipline, the benefits of reusing shared data have been asserted by a number of researchers, and data reuse has been a major concern in many disciplines. Assessing data for trustworthiness becomes important in data reuse with the growth in data creation because of the lack of standards for ensuring data quality and potential harm from using poor-quality data. This research explores many facets of data reusers' trust in data generated by other researchers focusing on the trust judgment process with influential factors that determine reusers' trust. The author took an interpretive qualitative approach by using in-depth semistructured interviews as the primary research method. The study results suggest different stages of trust development associated with the process of data reuse. Data reusers' trust may remain the same throughout their experiences, but it can also be formed, lost, declined, and recovered during their data reuse experiences. These various stages reflect the dynamic nature of trust.

Introduction

Data are used to generate new findings and are the basis of scientific research. Acquisition of the "right" data is significant in all research because inappropriate data (i.e., data that does not fit the research purpose or is of poor quality) may lead to distorted and unreliable results. Finding trustworthy data for reuse[1] is, thus, an important part of the research process. Faniel and Jacobsen (2010) believe that trustworthiness, the extent to which researchers can trust the

[1]Not many studies have formally defined the term *reuse*, but researchers generally understand it to indicate the use of data by someone who did not collect it. Therefore, reuse refers to a secondary use of data that is not defined by their original purpose but is intended to address new problems (Karasti & Baker, 2008; Zimmerman, 2008). Following this definition, this research defines *data reuse* as the secondary use of data by outsiders of the original studies.

data created by others, is one of the most important criteria for finding reusable data.

However, trust judgment is not a simple task for data reusers. McCall and Appelbaum (1991) point out that reusers must spend significant amounts of time absorbing information about data before deciding to use them (or deciding not to), as reusers are typically unfamiliar with the details of data that they have not collected. In addition, although other scholarly materials, such as journals or conference publications, have established systems to validate scholarly outcomes through peer-review processes, a validation or peer-review process for data has not yet been established as a norm in data-sharing and reuse, although discussions about these processes have emerged (e.g., Kratz & Strasser, 2015). This lack of standards for the trustworthiness of shared data is one of the difficulties in assessing data for reuse. Cultural institutions, including institutional or disciplinary repositories, are known as the places where "trusted" information is preserved, but communications and interactions around data are more dynamic than those surrounding traditional scholarly materials housed in repositories. Data reusers do not only rely on cultural and/or academic institutions to acquire materials such as data; they use interpersonal relationships to obtain the raw forms of information and data that are not processed, managed, or curated by professionals. In addition, although repositories are known to perform curatorial work for quality checks, such as correcting errors (Daniels et al., 2012), not all disciplines conduct data validation for methodological soundness before preserving data in trusted repositories, and few disciplines where large datasets are the norm (e.g., climate modeling and high energy physics) conduct thorough investigations of data quality and validity before preserving it (Adelman et al., 2010; Callaghan, 2015; Stockhause, Höck, Toussaint, & Lautenschlager, 2012).

Understanding the difficulty and complexity of judging trust is significant when trying to support data reuse. Given the importance of trust in data and the complexity of the landscape of data reuse, this research explores data reusers' processes of making trust judgments about data. Trust in data has received recent attention among data reuse researchers, exploring the factors that influence reusers' trust

in data (e.g., Birnholtz & Bietz, 2003; Faniel & Jacobsen, 2010; Van House, 2002; Wallis et al., 2007; Zimmerman, 2008) and in data repositories (e.g., Donaldson & Conway, 2015; Yakel, Faniel, Kriesberg, & Yoon, 2013; Yoon, 2014). Although previous research provides a foundational understanding of reusers' trust in data, this study focuses on the process of trust judgment and the dynamic nature of trust as it appears during this process. Emphasis on the judgment process through the theoretical lens of trust provides a rich understanding of data reusers' thoughts and perceptions beyond their behaviors. Capturing the dynamics of trust also provides valuable insights for data curation research by asking questions about what this dynamic means for data curation and what it suggests for curation practices.

Theoretical Framework

Trust Conceptualization

The concept of trust has been widely studied in various disciplines (e.g., sociology, social psychology, organizational behavior, marketing, and economics). Because researchers take varying approaches to understand the concept of trust through their own disciplines, a full consensus on the definition of trust has not yet been reached. However, a review of trust definitions reveals several commonalities. First, trust is associated with probable positive outcomes, such as one's belief that another party will perform actions that will result in positive outcomes for oneself (e.g., Anderson & Narus, 1990; Mayer, Davis, & Schoorman, 1995). This positive expectation is linked to risk-taking, which is one's willingness to rely on another's actions in a situation involving the risk of opportunism (Williams, 2001). Trust is also associated with confidence in the "other's good will" (Ring & Van de Ven, 1992). The confidence brings greater optimism about the decision as the risk in decision-making is less pronounced. The review also reveals that there is a relationship between predictability and trust, where trust is dependent on predictable behaviors (e.g. Zucker, 1986). These common concepts appear in the several most influential definitions of trust: "willingness to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor" (Mayer et al. 1995, p. 712); "a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another" (Rousseau, Sitkin, Burt, & Camerer, 1998, p. 395); "concerns a positive expectation regarding the behavior of somebody or something in a situation that entails risk to the trusting party" (Marsh & Dibben, 2003, p. 470).

One difference among these definitions is whether trust is seen as a belief or a behavior. Psychologists (e.g., Giffin, 1967; Good, 1988) conceptualize trust as a psychological trait or state that individuals develop in to degrees and, thus, see trust as a mental status or belief. Behavioral psychologists, on the other hand, propose a behavioral interpretation of trust by equating it to cooperation with others (Lewis & Weigert, 1985). Although these different perspectives could lead to fundamental differences in the understanding of trust, some studies do not clearly indicate which perspective they follow and some use mixed perspectives in their conceptualizations.

This study develops a working definition that reflects both approaches, that is, an understanding of trust in the context of data reuse in which it is considered both a psychological and a behavioral phenomenon: Trust is the data reusers' belief that the data will result in positive outcomes, leading to the reuse of such data in their research. Data reusers' trust judgments can be understood as psychological processes, and whether they accept and use certain data can be seen as an indication of trusting behavior.

Trust Development

Although early research on trust (e.g., economics) considers it to be static and social psychologists often see trust as an all-or-nothing concept (e.g., trustees either completely trust or distrust), other studies treat trust as something that can be changed, built, developed, and decreased by interactions or relationships (Rousseau et al., 1998). Several researchers (e.g., Doney & Cannon, 1997; Kelton, Fleischmann, & Wallace, 2008; Rousseau et al., 1998), who see trust as dynamic, describe the process whereby trust is developed. Rousseau et al. (1998) explains the three phases of trust—building (where trust is formed), stability (where trust already exists), and dissolution (where trust declines)—these phases of were limited to providing the simple status of trust existence.

A more developed version of the trust-building process is proposed by Doney and Cannon (1997) and Kelton et al. (2008). Doney and Cannon (1997) present a four-step trust-building process: prediction, capability, intentionality, and transference. Similarly, Kelton et al. (2008) provide a five-step trust development process: prediction, attribution, bonding, reputation, and identification. In Doney and Cannon's process (1997), prediction is based on the trustor's assessment of either the trustee's credibility and benevolence or the trustee's past behavior and promises. The capability step involves determining the trustee's ability to meet obligations, whereas the intentionality step involves the interpretation and assessment of the trustee's motives, out of which trust emerges. Kelton et al. (2008) use prediction more broadly, including Doney and Cannon's (1997) stages of prediction, capability, and intentionality, and explain that prediction is based on the past behavior of trustees. The attribution step refers to the assessment of the underlying quality or motivations of the trustees based on observations, just as the intentionality step does. The bonding stage proposed by Kelton et al. (2008) is the emotional development of a trustor–trustee relationship. The next stage is called reputation (Kelton et al., 2008), or transference (Doney & Cannon, 1997). This refers to the awarding of trust based on the recommendations of others. In this stage, trust is further developed and transferred to other parties as a "proof source." Kelton et al. (2008) add identification as the final stage,

which is developed when trustors and trustees share a common identity, goals, and values.

Although these previous studies provide a useful framework for understanding data reusers' trust development, these process do not well capture the fluctuation of trust (if it develops), such as what can happen after trust is developed and how fully developed trust can decline or be restored after its dissolution. Thus, this study does not start with a fixed model of trust development for analysis but adopts the conceptual similarities of trust development from previous studies during the later stages of its data analysis. Because trust development is not a one-time process for data reusers, new stages and processes in the context of data reuse are presented in the results section.

Influencing Factors on Data Reusers' Trust Assessment

Previous studies have identified several elements in reusers' trust assessments of data, whether the concept of trust was explicit or implicit in their discussion, including both the data's properties and their social aspects. Several researchers report that trust stems from factors inherent in the data, such as collection methods, measurements, or variables (e.g., Faniel & Jacobsen, 2010; Wallis et al., 2007), which are sometimes the primary source of trust (Zimmerman, 2008). Because assessing trust in data inevitably requires an in-depth understanding of its context (Jirotka et al., 2005), the importance of contextual information during data reuse is well recognized (e.g., Faniel, Kansa, Kansa, Barrera-Gomez, & Yakel, 2013), and the amount of information that can be accessed by any means is critical. Information can be obtained through reusers' previous knowledge; their familiarity with the artifacts and processes; or direct interaction with colleagues, experts, or data producers (e.g., Birnholtz & Bietz, 2003; Van House, 2002). Data reusers can also develop their trust by reviewing the documentation about the data (Faniel & Jacobsen, 2010; Faniel, Kriesberg, & Yakel, 2015). Knowing who created and collected the data helps to lessen reusers' concerns about data quality, and reusers consider the competence, commitment, and reputations of the data producers as important assessment criteria (Zimmerman, 2008). Data producers are also assessed by "communities of practice" (CoP) (Van House et al., 1998), which are "groups of people who share a concern, a set of problems, or a passion about a topic, and who deepen their knowledge and expertise in this area by interacting on an ongoing basis" (Wenger, McDermott, & Snyder, 2002, p. 4). Van House et al. (1998) argues that reusers tend to trust data from their CoP and ask, "Is he or she a part of our CoP?" or "Can he or she be trusted to have used accepted methods to collect, analyze, and interpret the data?" Reusers' trust judgments can be influenced by individual knowledge, skills, and experiences, all of which help not only to understand but also to judge data quality (Borgman, 2007; Zimmerman, 2008).

Previous research also shows that reusers' trust judgments can be closely related to the data repositories from which they get the data, particularly when they know about how the data have been processed and "cooked" (Carlson & Anderson, 2007; Yoon, 2014). Other organizational attributes of repositories, such as integrity, transparency, reputation, and structural assurance that guarantees preservation and sustainability, have also been identified as important trust factors (Yakel et al., 2013; Yoon, 2014), as have reusers' perceptions and awareness of the roles of repositories (Yoon, 2014).

Although this review of previous studies presents various factors that influence reusers' decision-making processes about the use of data and trust judgment, most studies that address trust formation do not address the detailed processes of making trust judgments and changes to trust after its formation. Thus, the author of this research investigates reusers' trust beyond trust formation and tracks those changes to trust that happen during the experiences of using data.

Research Method

As this research concerns individuals' perspectives and thoughts regarding their experiences, a qualitative research method was chosen.

Study Sample

To recruit a homogeneous sample, this study limited the data type to quantitative social science data and employed purposive sampling. Purposive sampling is one of the most important kinds of nonprobability sampling to identify relevant participants (Welman & Kruger, 1999) and is appropriate for qualitative research because of the depth of data that comes from the richness of the participants' experiences of the phenomena under investigation (Smith, 2004). To identify individuals who have had experience reusing data, the author used data citation tracking from major databases. Although discussions on data citation have recently emerged, standards or guidelines for citing data have not yet been fully established (Altman & King, 2007; Fear, 2013; Gray, Szalay, Thakar, Stoughton, & van den Berg, 2002; Mooney, 2011; Parsons, Duerr, & Minster, 2010). Although tracking data citations may have limitations (e.g., the exclusion of research that does not properly indicate data reuse), it is still an effective way to identify data reusers. The major databases were searched, including EBSCOHost, SAGE Journals, ProQuest Social Science, and ERIC. Data reusers were identified from a keyword search in the full text, using the search terms "secondary data" and "secondary analysis," which are more commonly used than "reuse" in the social sciences (Gleit & Graham, 1989; Hinds, Vogel, & Clarke-Steffen, 1997). The search was performed from October 2013 to January 2014, only for journal publications and conference proceedings published in the United States.

The initial search identified reusers in various disciplines, but this study chose public health and social work for three

reasons. First, in the search, social work and public health presented the largest number of researchers who performed quantitative data reuse; this empirical evidence shows that both disciplines have data-reuse cultures while providing enough potential study participants. Second, Guest and Namey (2014) argue that secondary data analysis plays a key role in modern public health research, and Sales, Lichtenwalter, and Fevola (2006) argue that the use of secondary data has been growing in recent years in social work, which becomes more active in federally funded research. Third, the disciplines share similar characteristics—both have a professional orientation and use several of same data sets—which was helpful in recruiting a homogeneous sample.

Data Collection

A total of 229 (public health: 123; social work: 106) potential participants who were identified from the search were contacted for an interview, and 58 researchers responded to the study's e-mail invitation, for a response rate of 25.3%. Among them, 38 responded affirmatively and were interviewed. Given the diverse geographic distribution of the study participants, phone interviews were conducted from May 2014 to September 2014. The duration of interviews varied from 40 to 95 minutes, and the average length was 60 minutes. There was no financial incentive for the participants.

Semistructured interviews were conducted following a predeveloped interview protocol. Semistructured interviews are nondirective in that the questions (usually open-ended) asked by researchers are used as triggers for further conversation (Willig, 2008). Nondirective, open-ended questions were asked about the participants' reuse experiences (e.g., the process of data discovery, the initial selection, the criteria for data reused, changes in their judgments about the data during the process of reusing, any other factors that influenced this process, and their thoughts and perceptions of data and trust).

The unit of analysis for this study is an individual and his or her data reuse experiences. The participants were encouraged to draw from their past experiences to answer interview questions, which were not necessarily limited to one single case of reuse—data reusers' trust can be developed based on their cumulative experiences, some of which may not be directly related to a single case.

Data Analysis

The interviews were recorded, fully transcribed, and analyzed using a qualitative data analysis tool, NVivo 10 for Mac. Several strategies were used to analyze the interview data, which facilitated the move from description to interpretation, from capturing initial thoughts to generating themes through iterative and inductive cycles (Smith, 2007). The author read the transcripts multiple times, wrote descriptive and exploratory notes, and coded inductively and openly using the labels that best described the participants' experiences and thoughts. The top labels used to categorize codes

during the initial analysis were developed by comparing and contrasting interview patterns, and the categorized themes were also developed based on the author's understanding of the big picture of research on data reuse and trust. Previous literature on trust development (e.g., Doney & Cannon, 1997; Kelton et al., 2008) influenced the final stages of data analysis, as the author found similarities with some concepts/stages of development. Two participants were contacted during the analysis for member checking to verify the experiences that they detailed during the interviews.

Research Participants

A total of 37 interview sessions were conducted with 38 participants (one session included two participants). The interviewees were researchers in various positions (PhD students, postdocs, assistants to full professors, and research scientists), with a mix of genders and ranging in age from their 20s to 70s. The interviewees' experience in the study disciplines ranged from a minimum of 2 years to a maximum of 45, with the average being about 15 years. The interviewees' experience using data in research ranged from two to 40 years, with the average being about 16 years.

Of the researchers in this study, 36 had used more than three different data sets as secondary data for their research. Several of the participants had obtained and reused research data from institutions, including federal and state government organizations, and several did so from individuals or individual research teams. Eleven participants had only reused data from institutions, seven participants had only reused data from individual researchers or research teams, and the remaining 20 had used both types of data for their research. Four participants sought information from data repositories in the process of acquiring data from both institutions and individual researchers.

Trust Development Throughout the Process of Data Reuse

The study participants' data reuse experiences consisted of several stages: searching and discovering, initial screening for relevance and ease of use, acquiring, investigating and exploring, and analyzing. Data reuse was not a linear process because the participants often conducted other activities at the same time (e.g., obtaining data files to evaluate the data). Throughout the process of data reuse, the participants employed a variety of strategies for making trust judgments and developing their trust. Initial trust development was a process of developing trust in the data before the participants directly interacted with or experienced the data, which was usually during the process of data discovery and initial screening. When the participants developed enough initial trust, they were motivated to move on to the next stage of provisional trust judgment. The provisional trust judgment usually involved a thorough understanding of the data, based on the participants' own experiences and explorations. The participants' trust could either develop or decline during this process. The final trust judgment was

based on a participant's response to a trust violation (if they experienced one), depending on whether the violation was successfully resolved and trust was restored. If there was no trust violation, participants' trust usually remained high till the end of their data reuse.

Initial Trust Development

The participants developed their initial trust during the stages of data discovery and initial screening based on relevancy and ease of use. Because the initial trust development occurred prior to the participants' own data exploration or reuse, many social elements surrounding the data affected their initial trust judgments. The process of initial trust development consisted of four different mechanisms: prediction, attribution, transference, and bonding. Often, these four mechanisms were interrelated and not mutually exclusive for the participants' trust judgments, as one instance of trust can be supported by multiple mechanisms. For instance, participants reported that the competence of the original investigators fostered trust, where both attribution and transference contributed to find competence, an element of trust; and integrity also fostered trust through the mechanisms of attribution and transference.

Although a low level of initial trust may not fully prevent the further investigation of data (as some reusers may still want to investigate the data despite their suspicions), initial trust played an important role in the early stages of data reuse in that it determined the likelihood of the reusers using the date in the next stage. Also, initial trust did not guarantee that reusers would ultimately trust the data until they had fully explored and used the data.

Prediction

Prediction was the process of developing initial trust based on past experiences using data. When participants had previously worked with the data, they already had formed trust in those data and were therefore willing to use them again. The participants who used the same data more than once said that "it was a natural process" (PS08) because "I already knew so much about it that it wasn't anything I really checked into too much further" (PP06). When data reusers found the data sources (original investigators) to be trustworthy from past experiences, they also tended to trust new data from these same sources based on the positive past experiences. PP01 said, "in the course of my work life, I've never had an issue with [original investigators'] data," which made PP01 trust their data.

Attribution

Trust was based on the data reusers' rational choices and judgments, stemming from what credible information was available concerning the data, which is through the process of attribution. The participants discussed three types of evidence that influenced their trust: the existing evaluations of the data, the competence of the original investigators, and the intentions and ethics of the original study that produced the data.

The existing evaluations. Evaluations of the data helped the participants to develop trust by seeking confirmation from already-trusted sources. For example, peer-reviewed publications were useful sources on which several participants relied. PS10 considered publications as "one step further to establish trust" in data. The fact that someone had published an article using the data also bolstered the participants' confidence that the data product was "acceptable through the normal channels of scholarly activities" (PP17). The number of times that the data had been used was another indicator, as data were "widely trusted by others" when the data were "widely used and widely cited," which is "evidence and a component of trust" (PP17). The fact that the original studies were funded by either government or nonprofit organizations can be an indication of peer-evaluation and communities' acknowledgement because it suggests the recognition of outstanding research. Participants perceived that data funded by organizations such as the National Institutes of Health (NIH) and the National Science Foundation (NSF) as "trustworthy in a sense" (PS02) because "[the study] [was] vetted by a group of their peers" (PS08).

The competence of the original investigators. The participants asked whether the original investigators were capable of generating quality data and could be trusted to use accepted methods to collect, analyze, and interpret those data. Although the general reputations of the original investigators can be another way of directly checking their competence (see Transference), the participants also searched for evidence of competence on their own by checking the original investigators' membership in a Community of Practice (CoP). Wenger et al. (2002) defines shared concerns, experiences, and practices as the characteristics of a CoP. In particular, participants wanted to make sure that the original investigators shared training similar to their own in collecting, analyzing, and interpreting data. PP14 said, "I look at educational training. So, do they have expertise in the areas that they're supposed to have." Training can be common in a discipline or subdiscipline; a few health researchers noted that they share a certain approach to data as "epidemiologists" (e.g., an emphasis on using national sample data). Training can also be interdisciplinary, such as in quantitative research methodology. PS03 said, "If I knew that the people who collected the data had no training like I had, [in] data collection and interpretation, (. . .) I'd assume that the questions were very poor and leading questions and so forth, or ambiguous questions." By checking the original investigators' home departments, the departments in which they had been trained, their main methodologies, and their research interests (which can imply a core methodology), the participants were able to ascertain whether the original investigators were part of their CoPs. As previous literature on CoPs argues (e.g., Hislop, 2004), there are already trust-based

relationships within a CoP that are enhanced by a consensual knowledge base, which helps the members to accept the opinions of others within the CoP.

The original studies' intentions and ethics. The participants also considered the original investigators' research ethics and integrity, meaning whether they had collected and managed the data ethically and refrained from manipulating it. However, the ethics of the original study and its investigators were not always easily verified or evident from the study itself; thus, the participants often used proxies in assessing research ethics. For instance, a study's funding sources can indicate issues such as conflicts of interests. As PP04 said, "If [a funder] is a pharmaceutical company, I would be very skeptical, even if they are very well known about using their data. [. . .] I would be very comfortable using [the] data, if it's funded by someone who doesn't have any vested interest."

When the participants dealt with the data produced by unfamiliar and unfunded individual researchers, checking their social identities as researchers was an option, as having the same social identity generates a sense of kinship and allows for foundational trust. There was a high propensity to trust members of a group with a shared identity, which is known as "in-group favoritism" (Aronson, Wilson, & Akert, 2010). The social identity discussed by the participants not only pertained to a specific CoP but also to more general research communities. PP16 talked about trust in individual researchers and their data, integrity, and good intentions "not to mislead anyone" after checking their social identities as professors. PS02 also attested to the academic integrity of the researchers: "Because they are researchers, (. . .) I assume that they are not even giving me that data [laughter] that they've made, they cooked up."

Transference

The process of transference refers to when trust is transferred from others. The participants developed their trust based on other people's perceptions of the data's trustworthiness, mainly through reputation and colleagues' recommendations. The participants had high expectations of data with a good reputation or that had come from reputable original investigators. PP17 acknowledged the importance of a parent study's reputation, saying, "The study is considered 'the' study." Furthermore, reputation implied some positive characteristics, such as rigor in research and the capability and integrity of the original investigators/research organizations—PS07 believed the "high level of integrity" of an original investigator because he was "world-famous." To PP04, the fact that the data had been collected by reputable organizations such as the Centers for Disease Control and Prevention (CDC) made "[me] lay the responsibility with that organization to collect the data properly (. . .) and as best quality."

Bonding

Bonding is the process of emotional attachment that develops from the relationship between data reusers and the parties relevant to the data. The participants developed affect-based trust from their emotional connections with the original investigators. Because bonding takes place before the data reusers have worked with the data, this affective trust was developed from interpersonal relationships and the cumulative experiences with the original investigators. These established relationships and social ties build a strong trust between people, which made the participants rely on parties whom they believe to be competent. PS13 admitted, "There wasn't really an objective evaluation of the quality," and there were "more subjective [aspects] like, 'we know that this person does good work, (. . .) and [so] it's probably fine.'" The participants tended to accept data that came directly from "the interpersonal connection" (PS13), "not some anonymous person" (PP10), because "familiarity breeds confidence" (PS12).

Provisional Trust Judgment

When the participants developed enough initial trust, they were motivated to acquire the data and start investigating them further. During the stage of provisional trust, the participants thoroughly examined different aspects and properties of the data from their own experiences to see if it met their expectations. From their experiences and investigations, the participants' trust was either strengthened (remaining as high as the initial trust or even solidified) or weakened. In an ideal case, PS09 noted, "Usually my level of trust goes up or it solidifies or it doesn't necessarily change a lot." However, sometimes the participants' trust decreased after they started using the data. As PS06 said, "Knowing more about the data [and] working extensively can increase the level of trust but not always;" in other words, "the honeymoon period [could be] over." PS08 echoed:

> PS08: There are plenty of data sets out there where it might be the case where there's this initial feeling that it's great, and then you start looking around, and you think, "Man, this is isn't so great. I'm not really sure.. . ."

These changes to the participants' trust judgments reflect the dynamic nature of trust.

First Impression

The participants discussed their positive or negative first impressions of the data, often from where they obtained the data (e.g., from sites where the participants downloaded the data or from data packages that consist of data files, documentation, and other information). PP03 recalled her first experience with data available from a website as "very nicely done and very nicely laid out," which gave her "a good feeling," so she assumed their "trustworthiness"

(PP03). After receiving and unzipping a compressed data package, PS09 recalled reservations that the package "was not cleaned and organized at all...[I] was like, 'uh-oh.'" PS09 took this first impression of the data as a warning, even before exploring the data further. In contrast, PS08 had a positive first impression of a data package, which signaled its trustworthiness:

> PS08: It was very well done, very professional. Everything was sort of in its place. It was clear that a lot of work and a lot of time had been spent putting all of [data package] together. This is a general indication to me that the data [are] more likely to be trustworthy than not, as opposed to something that's kinda thrown together, hodge-podge.

These interactions preceded a closer examination of the data files or documentation; that is, the participants began to judge trustworthiness only by looking at the data packages or data websites and forming impressions on how things were organized. This cognitive aspect of trust may not guarantee full trust in data, but it provided a cue for the participants' trust judgments. PS10 said, "you can picture [the data] in your head (...), what they actually look like," and her expectations were "the data [are] gonna look clean."

Intrinsic Properties of Data

The participants also discussed how several intrinsic properties of the data, such as validity, reliability, and scientific rigor, increased or decreased their trust. Because data validity proved that the data (variables and measures) were well-founded and accurately collected, the participants said good validity "made me feel confident and trust the data" (PP17). Because the participants were aware of difficulties in assessing different types of validity (e.g., construct validity) on their own ("in a way, you don't really know" [PP18]), they usually verified "a sort of face validity" by "looking at the data to see whether or not they make intuitive sense" (PS08) and checking "the ranges, the mins and maxes, [whether they] are believable for these data" (PP07). Similarly, the participants expressed their trust in data with good reliability, meaning the data would produce stable and consistent results across studies. PP18 said that initial trust increased further after exploring the data because "[the results] seemed consistent across the studies [that used the same data] that I looked [at], that said [the data set] is pretty trustworthy, (...) I believe it will do the same thing to me."

However, some of the participants reported cases in which their trust in the data decreased. PP07 found "a lot of unbelievable values" in the data and "start[ed] to get the sense that maybe the whole data set is messed up and it's not trustworthy." Sometimes the participants took extra steps to search for "trend[s] that we would expect based on what we know from other national data" and require that "the initial frequencies of the variables of interest have to make sense" (PP12) when the data were weighted to be nationally representative. PS19 conducted this comparison and found

mismatches, thinking "the data [set] is not right or it's not trustworthy."

The number of missing values can also influence the participants' trust. The participants understood that there could "always be missing responses" (PS02) in the data, whether from the survey respondents or because of mistakes on the part of the data collectors. However, too many missing data made the participants question the accuracy of the data and the results of the original study. PP13 had to "lower trust in data in our [team's] examination of the raw frequencies" because "they seem to be missing about 18% of [variable]. We're not sure if we can trust that data because that's a large portion of [variable] to miss."

A few participants also reported that their levels of trust decreased significantly because of serious errors or flaws in the data, which made them "concerned about the whole quality of the data" (PP07), that is, the "level of trust for the data went down" because of the flaws in data.

> PS07: As I started analyzing the data and getting kind of into the weeds, I realized that there was kind of a huge issue within this data. And it probably, it most likely had affected the outcomes of the original findings to a certain extent [...] although I had kind of gone through and looked at the sample size, a good funding source, reputable investigators...

Data Preparation and Management

Although not all of the participants mentioned the term *data management*, many discussed several management aspects and their influence on trust judgment. Most of the participants assumed that the original investigators would manage and prepare the data for reuse, which may or may not have been true.

Documentation was commonly discussed by all of the participants as a factor that enhanced (or diminished) trust. Although the importance of documentation in data reuse has been discussed by many researchers (e.g., Faniel & Jacobsen, 2010; Niu & Hedstrom, 2008), the participants said that good documentation can enhance the level of trust because they believed that documentation reflects the original study and investigators' characteristics: "You can tell from the documentation whether or not a research[er] was thorough and careful" (PS08). In addition, the participants acknowledged the efforts to prepare the documentation and said, "when [I] have stuff documented, that makes me feel more secure [...] because it means [...] you took the time [and] commit to it" (PS12). Documentation also provides some evidence of scientific rigor because it includes detailed information about methodology, measurements, and the data collectors themselves. Although the participants checked for validity and reliability themselves, reading about the appropriateness of the study design influenced their trust judgments: "There's a level of rigor that I attribute to that, that has been confirmed by what I've seen in the documentation" (PP01). Finally, the participants noted that when the

documentation presented information in a "very transparent manner" (PP15) and "nothing was hidden" regarding any decision or change that the original investigators made to the data (PS04), their levels of trust were increased. Interestingly, although insufficient documentation without key information hindered the process of reuse, it did not automatically make all the participants lose their trust in such data; PP01 noted, "I can't really tell [from] knowing nothing about [the] data [from a poor documentation] unless I have a full picture [of data]. It's not like I lost my trust, I can't make a judgment."

Sometimes the participants experienced several issues with the data caused by inappropriate management or mistakes in management. Inconsistency between the documentation and data files and the inconsistent use of variables or measures made the participants "really frustrated" (PS14). PS08 had experience working with inappropriate codes and labels, "3,000 observations of numbers that are meaningless," as did PP05, which "makes me trust the data [a] little less" as PP05 was not sure "what parts are wrong and what parts are correct."

Final Trust Judgment

Some of the participants who had steady, high levels of trust from their initial to provisional trust judgments moved on "to run the actual models to see what [was] happening" (PP05). Although it might be possible to run into other trust violations during the data analysis, the participants said that their trust did not change when they performed the final analyses, as "[they were] much more likely to take [their] time to really understand as much of the data as [they could] before [going] in and [doing data analysis], and if not, [they'd] be wasting a lot of time" (PS15). The participants who were confident that they had made the correct judgments about the data before analyzing them did not find any fault with the data.

On the other hand, the participants who had experienced trust violations and changed their levels of trust in data had to decide whether to accept or reject the trust violation, which led to either trust restoration or loss.

Restoration

The participants' levels of acceptance for trust violations varied depending on their need for the data and their judgments about the seriousness of the violations. Most of the participants, with a few exceptions (see Losing Trust), attempted to resolve the trust violations with external help. As the end-users of data, the participants felt that they were not able to solve the issues by themselves and reached out to the communities they considered helpful: First, they contacted the original investigators and, if this was not possible or productive, they contacted other community members.

For successful trust restoration, the participants had to find a proper justification for the trust violations. PS13 and PS15 ended up using data with documentation that was "just very poorly described" (PS13), although it hindered the process, because they found out that the poor documentation

was not relevant to the original investigators' intention of "hiding something," but "they just didn't document more than that" (PS15) because they hadn't considered the possibility of data sharing. Similarly, when PP02 approached the original investigators for a question of validity, the investigators gave clear answers and explanations about the data "without hesitance," which helped reduce PP02's suspicions. PS04 was also able to restore her trust because the original investigators "answered every single question that I had." Thus, the transparent and honest attitudes of the original investigators helped the process of trust restoration: PS12 said, "yes they make an error, but [...] the fact that they are willing to report that error builds my trust in them [and] you know the proper thing to do."

Although the role of the original investigators was important in providing justification and helping to resolve trust violations, other relevant communities were significant sources for attempts to restore trust. The other communities were usually sources that the participants already trusted, such as "people around me who are more advanced [in data and/or methodology]" (PP14) and data repository staff who the participants considered "people that had any energy and expertise about the data" (PS09). Listening to other reusers' experiences in working with same data and their suggestions on violations also helped the participants to feel more secure about using the data. PP03 noted, "We are kind of in the same situation," and other reusers "[would] understand why [he was] having this issue." PS04 noted that "talking among [other data reusers]" was really reassuring [for her] experience with [the] data" in resolving the violation and continuing to use the data.

Losing Trust

Unfortunately, trust restoration was not always successful for the participants, even though they said such experiences were relatively rare. The failure of trust restoration was because of one of two causes: the participants' attempts at restoration were unsuccessful or the trust violations from the provisional trust judgments were too serious for the participants to seek restoration. A few participants had negative interactions with the original investigators that did not justify the trust violations in that they only received "partial answers," which was "very frustrating" (PS09) or they had the impression that "[they] may not exactly know" (PP12). For example, PP16 found errors in the variables and after contacting the original investigators, she realized that "they only came to realize an error when it was brought up to their attention," which made her lose trust not only in the data but also the outcome of the original study.

One participant did not even try to restore his trust after finding serious flaws in the data during the provisional trust judgment period—PS07 believed that the errors within the data "likely had affected the outcomes to a certain extent." He did not move further and stopped using the data, saying:

> PS07: What had been reported, what had been presented and discussed were, kinda, the best view of the data. [I]n reality,

the data did have some problems that weren't apparent until you got deeply inside and started looking.

This experience influenced his trust judgment practices in that he not only checked out the original investigators, funding sources, and documentations, and so on but also "x amount of publications that had come from the data" (PS07).

Discussion

The results of this research reveal the social nature of trust, which align with several findings from previous data reuse studies. Although several social parties and communities known to influence data reuse, such as data producers, colleagues, and experts (e.g., Faniel et al., 2013; Van House et al., 1998; Zimmerman, 2008), affected data reusers' trust judgments in this study, other entities, such as data management and user communities, play unseen roles in data reusers' trust development. The data reusers developed their trust from multiple dimensions of associated entities rather than relying on one entity, and the multiple interactions and relationships among all relevant entities influenced trust formation and the development of trust judgments about the data.

Although most data reuse studies focus on trust formation and influencing factors and do not track changes after trust formation, the findings of this research demonstrate that trust can form, disappear, decline, be lost, and be recovered, which demonstrate the dynamic nature of trust in the context of data reuse. The mechanisms of initial trust formation in this study are conceptually similar to what previous studies on trust research suggest, including prediction, attribution, transference, and bonding (Chopra & Wallace, 2003; Doney & Cannon, 1997; Kelton et al., 2008). According to previous studies, the key process of trust formation is how the trusting entity can infer the trustworthiness of the entity being trusted, and data reusers usually infer the trustworthiness of data from the characteristics of the responsible parties. This step is carried out before building a direct relationship with the data, and these processes are performed until the data reusers perceive that the data are sufficiently competent and satisfactory, which increases their trust in them. Because data reusers develop trust judgments through a complex process, their initial trust does not always remain at the same level. Although the development of initial trust leads data reusers to have positive expectations for the data—which then encourages them to risk spending time investigating and using the data—there is also a possibility of trust violations. Final trust judgment depends on trust restoration, which may occur in light of the data reusers' responses to violations. Proper justification of the violation is key to the process of trust restoration, and trust restoration generally occurs when the trustor believes the violation to have been unintentional (Tomlinson & Mayer, 2009).

It is important to note that data reusers have different vigilance levels regarding violations. Although this study was not able to quantify the level of changes in data reusers' trust as a result of violations, the study participants did show different attitudes toward them. For example, some participants were willing to restore their trust in data with unintentional errors and mistakes or poorly documented data (even if it really hindered the reuse process); however, they did not accept serious violations, such as errors that the original investigators did not explain well or errors that directly impacted the data analysis. In addition, awareness of a violation did not always prompt a reaction: Some participants said such violations immediately reduced their trust, whereas others were willing to withhold judgment. These results suggest that data reusers may not change their trust attitudes or lower their trust based on a single violation and that not all trust attributes carry the same weight. Although this study did not investigate this aspect, there could be a relationship between reusers' needs and their levels of acceptance of violations.

The dynamics of trust judgment suggest several implications for data reuse practices and data curation research. Because a certain level of initial trust is an important starting point for data reusers, it is important to provide help for developing this initial trust. Although some initial trust mechanisms depend on individual researchers' personal experience (e.g., past experiences and interpersonal relationships with other researchers), strategies such as linking data with existing evaluations and author information, including professional identity and other scholarly records, can systematically support some other mechanisms. Trust violations during the provisional trust judgment are the most important consideration for data curation, as fewer violations increase the likelihood of data reuse. Well-curated data, from reusers' perspectives, can start with fewer errors in files and documentation and intuitive code names. More efforts are required to prepare and manage data, such as thorough documentation, and these efforts could ease the process of trust judgment by eliminating or minimizing the factors that cause breaches and violations. One good sign toward well-curated data would be an emerging discussion on data management practices, as the findings of this study include the identification of several data management activities that directly impacted the participants' trust in the data. Finally, providing mechanisms for trust restoration is another important consideration in data curation. Although data reusers have different tolerance levels for violations, most are willing to attempt to recover trust, as they understand the potential for human errors and mistakes in any research project. Restoration mechanisms help to deal with unexpected violations that are not caught during data curation. Communication appears to be a key to the process of restoration, as the data reusers seek out external help from either data experts or other reusers. Connecting reusers to the communities they try to reach would be a great start to support restoration process.

Conclusion

This research offers several contributions to the fields of data reuse and data curation research. Although the literature

on data reuse is expanding, few attempts have been made to explain data reusers' behaviors from a theoretical perspective. As theoretical perspectives often lead to deeper understanding, this research analyzes the thoughts, perceptions, and beliefs behind the behaviors and actions of data reusers. Trust is a useful theoretical concept to explore data reusers' behaviors and perceptions, and as explained in the introduction, current data reuse practices make the concept of trust even more relevant. The theoretical understanding developed in this research has the potential to be applied to other contexts of data reuse in future studies.

By drawing on a range of fields in which the concept of trust is important, including sociology, social psychology, economics, information systems, and organizational behavior, this study contributes to the understanding of the multiple facets of trust that are involved in data reuse. A data reuser's trust judgment is not a one-time, simple process. Various types and levels of trust interact to enable data reusers to make trust judgments regarding data.

As with any research, this study has some limitations. This study only dealt with one particular type of data—quantitative data—and the reusers of these data. Data reusers' experiences can vary depending on the data type, because of the different forms and formats of data, the methods used to acquire the data, and the process of understanding and analyzing the data. Thus, although this study contributes to the understanding of data reusers' trust judgments, the findings of this study may not be directly applicable to other types of data reusers. Another limitation of this study was the process used to identify potential study participants. Because the study participants were manually selected from major scholarly databases, there was a possibility that some authors in the databases did not use the terms "secondary data" or "secondary analysis." This may have inhibited access to other potential data reusers who could have been included in the study.

References

Adelman, J., Baak, M., Boelaert, N., D'Onofrio, M., Frost, J.A., Guyot, C., ... Wilson, M.G. (2010). ATLAS offline data quality monitoring. Journal of Physics: Conference Series, 219(4), 1–6. doi: http://doi.org/10.1088/1742-6596/219/4/042018

Altman, M., & King, G. (2007). A proposed standard for the scholarly citation of quantitative data. D-Lib Magazine, 13(3/4). Retrieved from http://www.dlib.org/dlib/march07/altman/03altman.html.

Anderson, J.C., & Narus, J.A. (1990). A model of distributor firm and manufacturer firm working partnerships. Journal of Marketing, 54(1), 42–58.

Aronson, E., Wilson, T.D., & Akert, R. (2010). Social psychology (7th ed.). New York: Prentice Hall.

Birnholtz, J.P., & Bietz, M. (2003). Data at work: Supporting sharing in science and engineering. In GROUP'03 Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work (pp. 339–348). New York, NY: ACM.

Borgman, C.L. (2007). Scholarship in the digital age: Information, infrastructure, and the Internet. Cambridge, MA: MIT Press.

Callaghan, S. (2015). Data without peer: Examples of data peer review in earth sciences. D-Lib Magazine, 21(1/2). doi: 10.1045/january 2015-callaghan

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. doi:10.1111/j.1083-6101.2007.00342.x

Chopra, K., & Wallace, W. (2003). Trust in electronic environments. Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03) 9(9), pp. 331–341.

Daniels, M., Faniel, I., Fear, K., & Yakel, E. (2012). Managing fixity and fluidity in data repositories. Proceedings of the 2012 iConference (pp. 279–286). New York, NY.

Donaldson, D.R., & Conway, P. (2015). User conceptions of trustworthiness for digital archival documents. Journal of the Association for Information Science and Technology, 66(12), 2427–2444. doi: 10.1002/asi.23330

Doney, P.M., & Cannon, J.P. (1997). An examination of the nature of trust in buyer-seller relationships. Journal of Marketing, 61(2), 35. doi:10.2307/1251829

Faniel, I.M., & Jacobsen, T.E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. Computer Supported Cooperative Work (CSCW), 19(3-4), 355–375. doi:10.1007/s10606-010-9117-8

Faniel, I.M., Kansa, E., Kansa, S.W., Barrera-Gomez, J., & Yakel, E. (2013). The challenges of digging data: A study of context in archaeological data reuse. In Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 295–304). New York, NY: JCDL '13. doi:10.1145/2467696.2467712.

Faniel, I.M., Kriesberg, A., & Yakel, E. (2015). Social scientists' satisfaction with data reuse. Journal of the Association for Information Science and Technology, 1–13. doi:10.1002/asi.23480.

Fear, K. (2013). Measuring and anticipating the impact of data reuse. (Doctoral dissertation). University of Michigan, Ann Arbor. Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/102481

Giffin, K. (1967). The contribution of studies of source credibility to a theory of interpersonal trust and the communication process. Psychological Bulletin, 68, 104−120.

Gleit, C., & Graham, B. (1989). Secondary data analysis: A valuable resource. Nursing Research, 38(6), 380–381. doi:10.1097/00006199-198911000-00018

Good, D. (1988). Individuals, interpersonal relations and trust. In D. Gambetta (Ed.), Trust: Making and breaking cooperative relations (pp. 31–48). Oxford, UK: Basil Blackwell.

Gray, J., Szalay, A.S., Thakar, A.R., Stoughton, C., & vandenBerg, J. (2002). Online scientific data curation, publication, and archiving. In the Proceedings of SPIE, 4846, 103–107. Bellingham, WA: SPIE. doi:10.1117/12.461524

Guest, G., & Namey, E.E. (2014). Public health research methods. Thousand Oaks, CA: SAGE publications.

Hinds, P.S., Vogel, R.J., & Clarke-Steffen, L. (1997). The possibilities and pitfalls of doing a secondary analysis of a qualitative data set. Qualitative Health Research, 7(3), 408–424. doi:10.1177/104973239700700306

Hislop, D. (2004). The paradox of communities of practice: Knowledge sharing between communities. In P.M. Hildreth & C. Kimble (Eds.), Knowledge networks: Innovation through communities of practice (pp. 36–45). Hershey, PA: Idea Group Publishing.

Jirotka, M., Procter, R., Hartswood, M., Slack, R., Simpson, A., Coopmans, C., & Voss, A. (2005). Collaboration and trust in healthcare innovation: The eDiaMoND case study. Computer Supported Cooperative Work, 14, 369–398. doi: 10.1007/s10606-005-9001-0

Karasti, H., & Baker, K.S. (2008). Digital data practices and the long term ecological research program growing global. International Journal of Digital Curation, 3(2), 42–58. doi:10.2218/ijdc.v3i2.57

Kelton, K., Fleischmann, K.R., & Wallace, W.A. (2008). Trust in digital information. Journal of the American Society for Information Science and Technology, 59(3), 363–374. doi:10.1002/asi.20722

Kratz, J.E., & Strasser, C. (2015). Researcher perspectives on publication and peer review of data. PLoS One, 10(2). doi: 10.1371/journal.pone.0117619

Lewis, J.D., & Weigert, A. (1985). Trust as a social reality. Social Forces, 63(4), 967–985.

Marsh, S., & Dibben, M. (2003). The role of trust in information science and technology. Annual Review of Information Science and Technology, 37(1), 465–498. doi: 10.1002/aris.1440370111

Mayer, R., Davis, J., & Schoorman, D. (1995). An integrative model of organizational trust. The Academy of Management Review, 20(3), 709–734. doi: 10.5465/AMR.1995.9508080335

McCall, R.B., & Appelbaum, M.I. (1991). Some issues of conducting secondary analyses. Developmental Psychology, 27(6), 911–917. doi: 10.1037/0012-1649.27.6.911

Mooney, H. (2011). Citing data sources in the social sciences: Do authors do it? Learned Publishing, 24(2), 99–108. doi:10.1087/20110204

Niu, J., & Hedstrom, M. (2008). Documentation evaluation model for social science data. Proceedings of the American Society for Information Science and Technology, 45(1), 11–11. Retrieved from hdl.handle.net/2027.42/63090

Parsons, M.A., Duerr, R., & Minster, J.-B. (2010). Data citation and peer review. Eos, Transactions American Geophysical Union, 91(34), 297–298. doi:10.1029/2010EO340001

Ring, P.S., & Van de Ven, A.H. (1992). Structuring cooperative relationships between organizations. Strategic Management Journal, 13(7), 483–498. doi:10.1002/smj.4250130702

Rousseau, D.M., Sitkin, S.B., Burt, R.S., & Camerer, C. (1998). Not so different after all: Across-discipline view of trust. Academy of Management Review, 23(3), 393–404. 10.5465/AMR.1998.926617

Sales, E., Lichtenwalter, S., & Fevola, A. (2006). Secondary analysis in social work research education: Past, present, and future promise. Journal of Social Work Education, 42(3), 543–560. doi: 10.5175/JSWE.2006.200404136

Smith, J.A. (2004). Reflecting on the development of interpretative phenomenological analysis and its contribution to qualitative research in psychology. Qualitative Research in Psychology, 1(1), 39–54. doi: 10.1191/1478088704qp004oa

Smith, J.A. (2007). Hermeneutics, human sciences and health: Linking theory and practice. International Journal of Qualitative Studies on Health and Well-Being, 2(1), 3–11. doi:10.1080/17482620601016120

Stockhause, M., Höck, H., Toussaint, F., & Lautenschlager, M. (2012). Quality assessment concept of the World Data Center for Climate and its application to CMIP5 data. Geoscientific Model Development, 5, 1023–1032. http://doi.org/10.5194/gmd-5-1023-2012

Tomlinson, E.C., & Mayer, R.C. (2009). The role of casual attribution dimensions in trust repair. The Academy of Management Review, 34(1), 85–104. doi: 10.5465/AMR.2009.35713291

Van House, N.A. (2002). Digital libraries and practices of trust: Networked biodiversity information. Social Epistemology, 16(1), 99–114.

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, WA.

Wallis, J.C., Borgman, C.L., Mayernik, M.S., Pepe, A., Ramanathan, N., & Hansen, M. (2007). Know thy sensor: Trust, data quality, and data integrity in scientific digital libraries. Proceedings of 11th European Conference, ECDL September 16–21. (pp. 380–391). Budapest, Hungary: Springer Berlin Heidelberg.

Welman, J.C., & Kruger, S.J. (1999). Research methodology for the business and administrative sciences. Johannesburg, South Africa: International Thompson.

Wenger, E., McDermott, R.A., & Snyder, W. (2002). Cultivating communities of practice. Watertown, MA: Harvard Business Press.

Williams, M. (2001). In whom we trust: Group membership as an affective context for trust development. The Academy of Management Review, 26(3), 377–396. doi:10.2307/259183

Willig, C. (2008). Introducing qualitative research in psychology (2nd ed.). Berkshire, UK: Open University Press.

Yakel, E., Faniel, I., Kriesberg, A., & Yoon, A. (2013). Trust in digital repositories. International Journal of Digital Curation, 8(1), 143–156. doi:10.2218/ijdc.v8i1.251

Yoon, A. (2014). End users' trust in data repositories: Definition and influences on trust development. Archival Science, 14(1), 17–34. doi: 10.1007/s10502-013-9207-8

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. doi:10.1177/0162243907306704

Zucker, L.G. (1986). The production of trust: Institutional sources of economic structure. Research in Original Behavior, 8, 55–111.