DATA REUSE AND USERS' TRUST JUDGMENTS: TOWARD TRUSTED DATA CURATION

Ayoung Yoon

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Ph.D. in the School of Information and Library Science

Chapel Hill
2015

Approved by:

Helen R. Tibbo

Cal Lee

Nancy McGovern

Barbara Wildemuth

Elizabeth Yakel

ABSTRACT

Ayoung Yoon: Data Reuse and Users' Trust Judgments: Toward Trusted Data Curation
(Under the direction of Dr. Helen R. Tibbo)

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 its 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 dissertation aims to explore many facets of data reusers' trust in data generated by other researchers, focusing on user-defined trust attributes and the judgment process with influential factors that determine these attributes. Because *trust* is a complex concept that is explored in multiple disciplines, this study developed a theoretical framework from an extensive literature review in the areas of sociology, social psychology, information, and information systems.

This study takes an interpretive qualitative approach by using in-depth semi-structured interviews as the primary research method. The study population comprises reusers of quantitative social science data from public health and social work—the primary disciplines with data reuse cultures. By employing purposive sampling, a total of 38 participants were recruited.

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. The user-defined trust attributes that influenced the formation of trust also suggested various implications for data curation.

The outcomes of this study will contribute to the current research on data reuse and data curation. Integrating theories and concepts of trust can provide a new theoretical lens to understand reusers' behaviors and perceptions. Understanding how data reusers trust data will also provide insights on how to improve current data curation activities in a user-trusted way, such as methods that ensure users' trustworthiness during data curation and develop user evaluation criteria for the trustworthiness of data.

ACKNOWLEDGEMENT

Every time I read the acknowledgment sections of other dissertations, I come across a phrase such as "this work would have never been possible without such and such's help." However, it wasn't until I came to write my own acknowledgment that I truly understood what this sentence really means. This dissertation would really never have been completed without countless people helping me to get through it. I should first thank all my participants, who gladly took the time to discuss their thoughts, lives, and research practices to support my research, despite their busy schedules and time constraints. Without their contributions, this dissertation would not have been possible.

I want to express my sincere gratitude to my advisor Dr. Helen Tibbo (chair) and other committee members. By offering her continuous support, encouragement, and the unrelenting honesty I needed to improve my work, Dr. Tibbo helped me to complete this dissertation. Her profound knowledge of research and of the profession in general proved invaluable to my progress. I'd like to thank Dr. Barbara Wildemuth, who gave me tremendous support and advice in terms of my preliminary studies and this dissertation. I'd also like to offer special thanks to Dr. Beth Yakel for her guidance and help since 2008 when I started my masters, and for the opportunity to work on the DIPIR project as a member. I am very grateful to Dr. Cal Lee and Dr. Nancy McGovern who have offered me comments, critiques, and suggestions that have helped me to improve my work.

I want to thank my doctoral colleagues at SILS. We supported each other when things got dark, not only by working together but also by enjoying our non-doctoral life together. In addition, the staff at SILS always gave me excellent administrative help while completing this dissertation.

I would like to express my sincere gratitude and love for my family. To my parents, Tae-won Yoon and Kyong-ja Choi, for their endless love, support, and trust, which lifted me up when I felt discouraged. To my beloved sister Hye-young Yoon, for her bright laughter, great humor, and ever-lasting friendship, which gave me joy when I needed it most (and of course, many surprise boxes from Korea!). And to my grandmother, Hee-jung Kim, who is now resting in heaven, for her love and belief in me, which I was not even fully aware of when I was young. To my parents-in-law, Dong-man Shin and Sook-ja Kim, who always encouraged and supported me through their love and prayer. Most importantly, I'd like to recognize the invisible and invaluable support and patience of my husband, Hosop Shin. A single "thank you" is not enough to express how much your doing this with me means to me.

Finally, I thank God for equipping me, through His Holy Spirit, with the strength, knowledge, skills, ability, and fortitude to complete this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1. Introduction and Statement of the Problem

1. Introduction and background

The definition of *data* varies by discipline, and data come in various formats and types. The National Research Council (1999) defines *data* as "facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors" (p. 15). The National Science Board (2005) uses the term *data* as "any information… including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc." (p. 13). The National Science Foundation classifies data into four types: (1) observational data (e.g., weather measurements and attitude surveys); (2) computational data (e.g., results from computer models and simulations); (3) experimental data (e.g., results from laboratory studies); and (4) records (e.g., from government, business, and public and private life) (Borgman, 2010, p. 19).

Despite the different definitions, there is an agreement that data sharing and reuse are beneficial in scientific research, and attention to data sharing and reuse is growing. Since the 1980s, researchers have been concerned about the issue of data sharing and reuse (e.g., Glaeser, 1990; Fienberg, Martin, & Straf, 1985)—though there was a difference depending on the field— and the discussions on sharing social science data and their value began even earlier, at the beginning of the 1960s (Clubb, Austin, Geda, & Traugott, 1985). Discussions of data sharing, reuse, and curation have been emerging for the past 10 years along with revolutions in the practice of science known as data-intensive research, or "eScience" (Kunze et al., 2011), and the

growth of data in "big science" (Borgman, 2009). Recently, awareness of the challenges and

needs for managing small science data has been developing (e.g., Carlson 2006; Cragin, Palmer,

Carlson, & Witt, 2010; Key Perspectives Ltd., 2010; Shorish, 2012), as well as issues regarding

qualitative data sharing and reuse (see more in *Qualitative Inquiry: Research, Archiving, and*

Reuse, 2005; *IASSIST Quarterly 34*(3&4), 2010).

While attention to data sharing and reuse is growing, controversies and challenges still

surround data-sharing practices. Commonly discussed issues and challenges include the

following: 1) no or little rewarding of professional credit (e.g., social recognition, tenure

promotion, data as a publication, etc.); 2) cost of sharing (e.g., time and effort to clean up data,

create documentation and metadata, and check the integrity and consistency of data); 3) concerns

about the misuse of data (e.g., the risk of data being incorrectly interpreted, combined

inappropriately, or incorrectly represented); 4) ethical issues (e.g., concerns about confidentiality

as certain data may contain sensitive personal information); and 5) researchers' reluctance due to

the researchers' keen sense of ownership of their data (Baker & Yarmey, 2008; Borgman,

Wallace, & Enyedy, 2007; Campbell et al., 2002; Hilgartner & Brandt-Rauf, 1994; Louis, Jones,

& Campbell, 2002; Niu & Hedstrom, 2008; RIN, 2008; Sieber, 1991; Van House, 2003).

Despite the challenges and barriers, data sharing and reuse have been on the rise, with

efforts to overcome the barriers like data publications (e.g., Costello, 2009; Lawrence, Jones,

Matthews, Pepler, & Callaghan, 2011; Smith, 2009), and with enforcement by funding agencies

(e.g., Wellcome Trust Statement on Genome Data Release, 1997; Wellcome Trust Policy on

Access to Bioinformatics Resources by Trust-funded Researchers, 2001; Sharing Data from

Large-scale Biological Research Projects: A System of Tripartite Responsibility, 2003; ESRC

Research Data Policy, 2010; National Institutes of Health (NIH)'s data-management plan

requirement in 2003; National Science Foundation (NSF)'s new data-management requirements in 2010). Data repositories, such as the Protein Data Bank and the Inter-university Consortium for Political and Social Research (ICPSR), also serve as resources for data sharing and reuse while promoting sharing practices among scientists.

The benefits of reusing shared data have also been asserted by a number of researchers. Birnholtz and Bietz (2003) and Borgman (2011) argued shared data can be used not only to validate existing results but also to generate new findings built on the work of others. Re-analyzed data or data combined with new data can also help to verify published results or arrive at new conclusions (National Academy of Science, 2009). Thus, research data must be available for use beyond the purposes for which they were initially collected to enable others to ask new questions of extant data, advance solutions for complex human problems and the state of science, reproduce research, and expand the instruments and products of research to new communities (Borgman, 2010; Borgman, 2011; Hey & Trefethen, 2003; Hey, Tansley, & Tolle, 2009).

Data sharing and reuse are not yet be the norm in every discipline. However, much research has been recently conducted in the area of recognizing the benefits of sharing and reusing data. One key issue is addressing and meeting end-users' current and future needs, as the fundamental purpose of sharing data is for them to be used by other researchers.

2. Problem statement

2.1. Data reuse and reusers' trust judgment

Data play a vital role in research because they are used to generate new findings and are the basis of scientific research. Researchers conceive the value of data as a way to address

specific gaps in knowledge, and valuable data help researchers to answer their research questions and/or test hypotheses (Akmon, 2014). Thus, acquisition of the "right" data is significant in all research because inappropriate data (e.g., data that do not fit the research purpose, poor-quality data, and so on) may lead to distorted and unreliable results. For data reusers, the process of acquiring appropriate data for research is different than for researchers who collect their own data. Reusers need to select reusable and trustworthy data based on given information about the data.

A previous study identified three criteria of reusability of data: relevancy, understandability, and trustworthiness (Faniel & Jacobsen, 2010). Relevancy is the degree to which data answer research questions; understandability refers to whether the intended meaning of data can be easily perceived; and trustworthiness is the extent to which researchers can trust data created by others, such as others in different disciplines. While relevancy, usability, and trustworthiness are the key elements affecting data reusability, assessing data and finding trustworthy data for reuse are important parts of the research process, especially given the value and role of data in research.

However, deciding whether data reusers can trust data or not is not a simple job. One reason for this is the lack of standards for the trustworthiness of data during the process of data sharing. While other scholarly materials, such as journals or conference publications, have established systems to validate scholarly outcomes through peer review, validation or peer review of 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). In addition, many disciplines, including the social sciences, do not conduct data validation or quality checking before preserving data in a trusted repository and only assess data when data are used by other

people or when data are submitted to a data journal (Callaghan, 2015). Furthermore, few disciplines where large data sets are the norm (e.g., climate modeling and high energy physics) conduct thorough investigations on data quality before preserving data in a trusted repository (Adelman et al., 2010; Callaghan, 2015; Stockhause, Höck, Toussaint, & Lautenschlager, 2012). At the same time, communications and interactions around data are more dynamic than those which surround traditional scholarly materials. Researchers do not just rely on cultural and/or academic institutions to acquire scholarly materials, including data, but also make use of interpersonal relationships to obtain raw forms of information and data that are not processed, managed, or curated by professionals. Lastly, because reusers are typically unfamiliar with the details of data they have not created, the trust they place in data is often dependent on individual knowledge and experience. As McCall and Appelbaum (1991) pointed out, this means that reusers must spend significant amounts of time absorbing information about data before using them (or deciding to use them).

Previous literature suggests that uncertainty and risk are the two conditions relevant to the concept of trust, and these conditions are apparent in data reuse. Data reusers' uncertainty about the data they use and the risk of using inappropriate data for research indicate the need for understanding data reuse and users' experiences through the theoretical lens of trust.

Many researchers have argued that trust is fundamental in human relationships and in society (e.g., Gambetta, 1988; Weber, Malhotra, & Murnighan, 2004), and see trust as an enabler of cooperative behavior (Gambetta, 1988); trust promotes adaptive network relations (Miles & Snow, 1992) and reduces harmful conflict (Meyerson, Weick, & Kramer, 1996). This fundamental significance of trust in society suggests that trust will also play a bigger role in the context of data reuse, where various types of relations and communications are involved among

data producers, data curators, data reuser communities, and other research communities. Thus, assessing reusers' trust in data is a complex process due to the various social, organizational, and individual factors that influence assessment.

2.2. Data reuse, data curation, and end-users' trust

Previous research has demonstrated the relationship between data curation and data reuse and has suggested that well-curated data is an integral part of data reuse. Coates (2014) argued that, because data are a key piece of the scholarly record, the management of data has an impact on the integrity of the scholarly record and on the potential for data sharing and reuse. Steinhart et al. (2008) argued that a well-developed data curation infrastructure, by exposing data for reuse, would enable new discoveries and ensure access to and preservation of scholarly outputs. The Digital Curation Centre (DCC) (n.d.) also argued that good practices of data curation can support data reuse in multiple ways; they ensure that the appropriate steps are taken to make data available in the first place (i.e., by presenting data and their associated descriptions in forms that are accessible and understandable to reusers); they prevent the unauthorized use of data (i.e., by maintaining legal constraints and usage rights); they provide the means of assuring data integrity and authenticity; and they enable reusers to be able to access high-quality data they can trust.

Given that one of the fundamental purposes of data curation is to support current and future use, it is important to consider what data curation means to data reusers and how it can help them. Understanding and meeting reusers' needs and expectations is important to enhance data reusability because curators decide what information to collect, provide, and preserve based on reusers' needs and expectations. Trust is a useful concept to understand users' expectations and needs, as the concept of trust is woven into the lifecycle of data—from the creation,

preparation, and management of data to their sharing and reuse to their preservation—and into the relations with parties involved in this lifecycle.

Trust is not a new concept in the field of archives, which traditionally is responsible for the curation of information. Speck (2010) said the concept of trust has been considered an integral component in the existence of archives, which made people expect a large volume of scholarly literature to be produced on the subject. However, Speck (2010) argued that discussions of trust have been limited either to discussions related to the ethics of the archival professions (e.g., Dingwall, 2004) or to the notion of "trusted" digital information and repositories. While archival and curation communities have understood the term *trust* as a synonym for "reliable" and "authentic" in relation to curation activities (RLG/OCLC, 2002, p. 8), little research exists on how (potential) users perceive the concept of trust in the context of curation.

Recently, several research studies have investigated users' trust in repositories where curation activities are conducted for digital information (as well as data) (e.g., Yakel, Faniel, Kriesberg, & Yoon, 2013; Yoon, 2014), and in documents within those repositories (Donaldson, 2015). These studies have broadened the understanding of users' perceptions of trust in data curation. However, understanding users' trust in data and data curation requires broader perspectives and context than simply trust in repositories. The patterns of searching and consuming data are diverse, and data reusers do not always acquire data through repositories. At the same time, different stakeholders are engaged in conducting data curation activities, including data producers through data management planning. This action positions data reusers in a different context and suggests the possibility that reusers' trust is not just tied to institutions that traditionally conduct curation activities. Investigating data reusers' perspectives on trust is

essential in this regard and will provide valuable insights into data curation research and the curation of data in a user-trusted way.

3. Purpose of the study and significance

In light of the importance of trust in the context of data reuse as demonstrated in the previous section, the proposed study aims to further explore the many facets of data reusers' trust, focusing on user-defined trust attributes and the judgment process with influential factors that determine these properties. Specifically, the aims of this study are as follows: (1) to understand data reusers' process of reuse; (2) to explore the reusers' process of developing trust during data reuse with factors that are influential in this process; and (3) to investigate the nature of trust attributes and characteristics that are defined by data reusers.

The outcomes of this dissertation study will contribute to the current research on data reuse and data curation in several ways: in theory, research, and practice. From a theoretical perspective, integrating theories and concepts of trust can provide a new theoretical lens to understand data reusers' behaviors and perceptions. The theoretical framework can offer insight and a deep understanding of individual researchers' beliefs, attitudes, thoughts, and perceptions beyond their behaviors of data reuse. Thus, the theoretical analysis will provide rich understanding, more than simply describing data reusers' practices and behaviors.

Second, this study will contribute to different domains of research. The study will enhance the understanding of the concept of *trust* by highlighting various facets of trust in the data reuse context through empirical research. Although previous research has reported assessing trustworthiness of data is an important part of data reuse processes, less research has been conducted regarding what trust really means to data reusers and how trust is formed or

developed. At the same time, full agreement on the meaning of trust has not yet been reached across various disciplines and in the data curation community, possibly due to the complexity of this concept. Thus, exploring data reusers' trust will add a valuable discussion to the current trust research.

This study can also provide valuable insights into the domain of scholarly communication and data curation. Reusing data requires researchers to have a deep understanding of these data and the different types of communications and interactions with relevant parties, which are often engaged during the process of understanding. By disclosing scientific communications about data for reuse, this study will contribute to the research area of scholarly communication. In addition, understanding data reuse practices and reusers' perspectives on data is important in data curation. In data reuse, it is significant to understand data reusers' needs and expectations and prepare data in a way that meets these needs and expectations. Data curation research, thus, will benefit by understanding users' perspectives and bringing them back to the current curation research. Knowing how data reusers trust data will provide insights on how to manage data in a user-trusted way.

Finally, from a practical perspective, this research will help to develop useful guidelines, recommendations, and standards for improving current data curation activities, such as methods to ensure trustworthiness of data during the data curation lifecycle and user evaluation criteria for trustworthiness of data. While the research findings cannot be directly generalized for several reasons (e.g., samplings and interpretive research, see the section regarding limitations in Chapter 6), this study can also provide insights and a preliminary understanding for developing trusted curation practices and standards across different data reuse contexts from reusers' points of view by providing the basis for a comparative study in the future.

4. Proposition of the study

This study presents two propositions regarding data reuse and the concept of trust. As there are different approaches to understanding data, data reuse, and trust, it will be useful for this study if I explain my understanding of these concepts based on past research.

4.1. Understanding data reuse as knowledge reuse

Several studies on data and knowledge imply a potential relationship between data reuse and knowledge, and these provide a new approach to understanding data reuse as knowledge reuse. For instance, Markus (2001) articulated data reuse as one of the most prominent examples of reusing knowledge, and Birnholtz and Bietz (2003) saw the process of understanding data for reuse as a process of knowledge transfer. According to Gold (2013), using data is a part of a knowledge process because, "unlike published textual narratives, data by its very nature lends itself to being moved around, filtered, added to, visualized, and linked with other data as part of knowledge design process" (p. 5). Unlike the traditional approach that defined the characteristics of data as explicit and objective and understood data as a product and fact (Galliers & Newell, 2003; Gold, 2013), this new approach views data as not only a product but also a process. Gold (2013) offered the following explanation:

> As the flow of data through their life cycles is studied, it has become more clear that digital data is not "objective" in any transcendent or permanent sense. Rather it is the outcome of agreements and decisions that have been used (with a range of skill and accuracy) to imagine and execute the use of instruments and practices that identify, gather, and record data. (p. 5)

When data are viewed as a process, reusing data becomes a process of interpretation. As Birnholtz and Bietz (2003) argued, in knowledge transfer, documentation is inherently insufficient and thus, no matter how well the data have been documented, they will always remain an incomplete representation of decisions and choices (Gold, 2013). Thus, the process of "imagination of data itself," as Gitelman (2013) referred to it, is necessary to understand data; "every discipline and disciplinary institution has its own norms and standards for the imagination of data, just as every field had its accepted methodologies and its evolved structures of practices" (p. 3).

This approach, which sees data as not a fixed fact but as a process, also aligns with current discourse on data curation that can help the process of "imagination of data itself" for both current and future data reusers. Angevaare (2009) argued that curating research data does not simply mean curating sources but also curating "knowledge," and Gold (2013) followed by saying, "the work of curation is a form of participation in the process of knowledge" (p. 8) because the purpose of data curation is to enable current and future use, which in turn makes it possible to produce new questions from the managed data (Gold, 2013).

Linking data and knowledge in a data reuse context suggests even greater relevance for trust and its roles. Previous research demonstrated the importance of trust in a knowledge-sharing environment. For instance, trust has served as a mediating factor that influences knowledge sharing and as a factor in the enhancement of knowledge sharing behaviors (e.g., Ho, Kuo, Lin, & Lin, 2010; Renzl, 2008). The role of trust in communities, such as *communities of practice* and *epistemic communities*, in relation to data reuse, has also been argued. Thus, this research will discuss data reusers' trust in more fruitful ways by adopting the approach of knowledge reuse and curation.

4.2. Understanding trust as a psychological state and a behavioral indicator

Previous studies on trust take various approaches to understand the concept of trust (see Chapter 3). From my extensive literature review of previous research on trust, I have developed a working definition of *trust* that reflects my approach and understanding of trust in the context of data reuse. The definition was never given to the participants of this study, as they were encouraged to develop their own understanding of trust based on their experiences and using their own words. As a working definition, I defined *trust* as data reusers' belief that data will result positive outcomes, leading to the reuse of data in their research. While previous studies defined *trust* either as a belief (a mental psychological status) or a behavior (a behavioral interpretation of trust), I adopted a mixed approach in which trust is considered both a psychological and a behavioral phenomenon. Data reusers' trust judgments can be understood as psychological processes occurring in their minds, and whether they accept and use a certain dataset can be seen as an indication of trusting behavior.

5. Structure of the dissertation

This dissertation consists of six chapters. Chapter 1 has explained the changes in scientific scholarship from emerging discussions on data sharing and reuse. These changes and the growing need for data reuse, as well as the role of data curation to support data reuse, serve as the motivation for this study. The usefulness of employing the concept of trust, which is the overarching construct of this dissertation, to understand data reusers' perceptions and behavior, has been discussed. Chapter 2 reviews the relevant literature in the area of data reuse and data curation and their connection to trust. Chapter 3 explores the concept of trust from the extensive studies previously conducted in this area, including studies from sociology, social psychology,

economics, organizational behaviors, and information systems research. Because trust is a

complex concept, understanding the varying approaches to trust is necessary to investigate its

multidimensional nature. Chapter 4 introduces the methodological stance of this dissertation and

interpretative qualitative research, and it describes the main method for data collection and

analysis. Chapter 5 reports and discusses the research results along with important points

emerging from the data analysis. Lastly, Chapter 6 summarizes the research findings and the

contributions of the dissertation, and concludes with the limitations of this study and directions

for future studies.

Chapter 2. Literature Review

1. Data reuse and trust

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). Broadly, reuse includes the reproduction or replication of prior study results as it contributes to the existing knowledge (King, 1995). Recently, the concept of repurposing has been added to the discussion of data reuse. In this context, data reuse has been defined as the use of data more than once for the same purpose, while data repurposing has been described as the use of data for a completely different purpose (Data Governance and Quality, 2012). Faniel and Jacobsen (2010) pointed out that the absence of a reuse definition causes major challenges in providing reusable data, even though other studies have demonstrated that data reuse can be beneficial to researchers.

1.1. Challenges of data reuse

1.1.1. Transferring context information of data

A number of studies have shown the contextual nature of data from the examination of data reuse practices and also by describing data reusers who search for context information from sources including journals and colleagues (e.g., Berg & Goorman, 1999; Bishop, 1999; Sandusky & Tenopir, 2007; Stewart, 1996; Zimmerman, 2007). Drawing on their work in earthquake

engineering, HIV/AIDS research, and space physics, Birnholtz and Bietz (2003) suggested, "data are not simple carriers of meaning," and noted that, "converting raw data into scientific or social meaning is an active, context-dependent process" (p. 341).

While data-context information is fundamentally important to understanding data for reuse, transferring contextual information is not a simple process; data cannot be simply handed off from one research project to another because comprehending data is not simple. Data are originally created for a specific purpose; a local context is therefore embedded in the data. For data to be reused beyond the original purpose, they must be moved from the local sphere to the broader world (e.g., for interdisciplinary research), which creates a distance from the original context (Zimmerman, 2008). This distance creates problems when reusers try to understand how data were used (e.g., Berg & Goorman, 1999; Cragin & Shankar, 2006; Faniel & Zimmerman, 2011; Jirotka et al., 2005; Zimmerman, 2008).

In addition, researchers have often argued about the difficulty of deciding what contextual information is important for reusers beyond the original research purpose (Birnholtz & Bietz, 2003; Carlson & Anderson, 2007; Markus, 2001). As one way to identify contextual information, Baker and Yarmey (2008) described the physical environment in which the data were produced, and the technical and social environments associated with obtaining data. Chin and Lansing (2004) listed more specific scientific and social contexts in which data were created, including information about physical, technical, and social environments (Table 1). Because this list was created in the context of scientific-experiment data, it may not apply to all data types across disciplines, and further research must be conducted in different disciplines.

Lastly, each reuser has different technical skills and tacit knowledge about understanding

data from other researchers and other fields, which creates another layer of difficulties of

understanding (Faniel & Zimmerman, 2011).

Table 1. *Scientific and social contexts (Chin & Lansing, 2004)*

Context information	Definition
General dataset properties	Owner, creation date, size, format, etc.
Experimental properties	Conditions and properties of the scientific experiment that generated or was applied to the data
Data provenance	Relationship of data to previous versions and other data sources
Integration	Relationship of data subsets within a full data set
Analysis and interpretation	Notes, experiences, interpretations, and knowledge generated from the analysis of data
Physical organization	Mapping of data sets to physical storage structures such as a file system, database, or some other data repository
Project organization	Mapping of data sets to project hierarchy or organization
Scientific organization	Mapping of data sets to some scientific classification, hierarchy, or organization
Task	Research task(s) that generated or applied the data set
Experimental process	Relationship of data and tasks to overall experimental process
User community	Application of data sets to different organizations of users

1.1.2. Documentation for data reuse

As previously shown, to enable data reuse, contextual information and knowledge about

the data must transfer from data producers to data reusers. As Niu and Hedstrom (2008) argued,

documentation is one channel for this information. Niu and Hedstrom (2008) defined

documentation as materials accompanying data that provide information about the data, and they

explained that it usually includes codebooks, related bibliographies, data-collection instruments,

and metadata for resource discovery.

There is a disagreement regarding whether the documentation data producers create for

themselves can be useful to others; some argue that it has limited use to others (e.g., Birnholtz & Bietz, 2003; Shankar, 2007; Zimmerman, 2008), while others have demonstrated that it is useful to reusers (Carlson & Anderson, 2007; Faniel & Jacobsen, 2010). Well-prepared and well-described documentation, however, enhances reusers' understanding. Niu and Hedstrom (2008) explained that documentation helps users decide whether data matches their research interests—in other words, they aid in assessing relevancy.

Despite the importance of documentation, a number of researchers have reported problems and challenges with it (e.g., Borgman, 2007, Corti, 2000; Fienberg, Martin, & Straf, 1985; McCall & Applebaum, 1991; Niu & Hedstrom, 2008; Van den Berg, 2005; Zimmerman, 2003). Similar to the challenges of transferring contextual information embedded in data, documentation can be inherently insufficient. Birnholtz and Bietz (2003) pointed out the innate challenges of documentation by observing, "Knowledge transfer in this instance is not simply a matter of sharing a set of instructions, but is a highly social process of learning practices that are not easily documented" (p. 341).

Borrowing the theory of communication reductionism, Niu (2009) explained that documentation is inherently inadequate because of the nature of tacit knowledge and communication reduction. Because tacit knowledge exists in cognizance of the human mind, it is difficult to formalize. People sometimes know things implicitly. In addition, due to the nature of communication, it is inevitable to miss information; not everything can be transferred (Carlson & Anderson, 2007; Niu & Hedstrom, 2008; Niu, 2009). Data producers' tacit knowledge or unconscious memories or information about data cannot be documented, even though there are core pieces of information.

Another problem is the issue of poor documentation. As discussed above, one of the challenges of data sharing, from the perspectives of data producers, is that the time and effort to create documentation is substantial (Baker & Yarmey, 2008). Data producers might not want to prepare documentation well because of a lack of motivation or because they do not have the skills to do so. Poor documentation can also result from the misunderstanding of reusers' needs; data producers cannot adequately document the data if they are not aware of those needs. For reusers, insufficient documentation causes a major problem because they either misunderstand the data or fail to understand them fully.

Some of the challenges of data reuse potentially influence reusers' trust judgments as they hinder acquiring a full picture of the data with all necessary information. Issues such as lost information during knowledge transfer are inherent and may not resolve instantly. However, other issues, such as identifying context information and documentation can be improved by good data curation practices and further research in this area.

1.2. Influencing factors on data reusers' trust assessment

Previous studies have identified and discussed several elements in relation to reusers' trust assessments. Reusers' trust judgments are not only related to the properties innate to data (e.g., technical aspects of data), but also to social aspects. In her study on ecologists, Zimmerman (2008) found there are social elements in all aspects of data reusers' experiences. Another element that plays a role in reusers' trust judgments is whether the data come from a repository.

1.2.1. Communities of Practice (CoP)

One major source of trust for data reusers identified in the literature is "Communities of

Practice" (CoP) (Van House, Butler, & Schiff, 1998). The term "CoP" was first introduced by

Lave and Wenger (1991), and is defined by three characteristics: a joint enterprise, mutual

engagement, and shared repertoire (Wenger, 1998). Though Lave and Wenger (1991) initially

developed this concept where learning took place through the process of legitimate peripheral

participation, Wenger, McDermott, and Snyder (2002) later redefined CoP as "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" (p. 4).

This definition has been applied to different communities, not just joint enterprises but

also to other occupational or professional communities, even if geographically distributed (Cox,

2005). Groups that are not directly or geographically connected to each other but still engage in

similar activities and share knowledge are called a "network of practice" (Brown & Duguid,

2001), which can be seen as a part of CoPs. There can be different types of CoPs among

researchers, including the data reuser group itself (e.g., within laboratory, disciplines, interest

groups, and more), whether they are physically connected or geographically distributed as a

group that shares practices, experiences, understandings, technology, and languages. As

members of the group, they learn not only practices of communities but also other members'

views and understanding to share a way of the world (Lave & Wenger, 1991).

Previous studies on CoPs have suggested that CoPs help to share knowledge based on

trust-based relations, which are enhanced by a consensual knowledge base and shared identity

(e.g., Hislop, 2004). Considering data not just as a source but as knowledge can help CoPs of

data reusers form trust within their CoPs. Specifically, they can judge whether they can trust data

depending on whether the data are generated by their CoP. For instance when looking at data

producers, the following questions should be asked: "Is he or she a part of our CoP?" "Can he or

she be trusted to have used accepted methods to collect, analyze, and interpret the data?" "Do we speak the same language?" "Do they see the world the same way that we do?" (Van House et al., 1998).

1.2.2. Epistemic communities

"Knowledge communities" or "epistemic communities" are another major source of trust. Haas (1992) defined epistemic communities as "a network of professionals with recognized expertise and competence in a particular domain and an authoritative claim to policy-relevant knowledge within that domain or issue-area" (p. 3) with four characteristics:

> (1) A shared set of normative and principled beliefs, which provide a value-based rationale for the social action of community members
>
> (2) Shared causal beliefs, which are derived from their analysis of practices leading or contributing to a central set of problems in their domain and which then serve as the basis for elucidating the multiples linkages between possible policy actions and desired outcomes
>
> (3) Shared notions of validity—that is, intersubjective, internally defined criteria for weighting and validating knowledge in the domain of their expertise; and
>
> (4) A common policy enterprise—that is, a set of common practices associated with a set of problems to which their professional competence is directed, presumably out of the conviction that human welfare will be enhanced as a consequence. (Haas, 1992, p. 3)

In addition, Haas (1992) explained that not only natural scientists but also social scientists or individuals from any discipline or profession can form epistemic

communities when they have "a sufficiently strong claim to a body of knowledge that is valued by society" (p. 16). However, epistemic communities often differ from the broader scientific communities or disciplines that share a set of casual approaches and knowledge but lack the shared normative commitments of members of epistemic communities. Because epistemic communities' ethical standards arise from their principled approaches to the issues at hand, their beliefs or goals can be different from the broader body of communities (e.g. economists as a disciplinary community vs. Keynesians as an epistemic community) (Haas, 1992, p. 19).

Researchers have often compared this notion of epistemic communities to CoPs, but others have argued that the role assumed by an epistemic community is distinct from a CoP as discussed by Brown and Duguid (1991) and Wenger (1998), although there are overlaps (Arena, Lazaric, & Lorenz, 2006; Lazaric, 2003; Lorenz-Meyer, 2009; Meyer & Molyneux-Hodgson, 2010). While CoPs tend to be more informal and focus on "non-intentional knowledge work," knowledge exchange between experts in epistemic communities is enabled by the codification of tacit knowledge (Lorenz-Meyer, 2009). In addition, epistemic communities' cognitive functions are not limited to the exchange of tacit knowledge. With substantial authority, epistemic communities extend their roles to the validation (or invalidation) of a particular practice and dissemination of explicit knowledge (Arena, Lazaric, & Lorenz, 2006; Lazaric, 2003).

Previous research on social epistemology has suggested that trust is related to members in epistemic communities taking strong or weak social views on the role of communities as holders of knowledge (Faulkner, 2010; Poutanen, 2001). As Van House (2002) suggested, epistemic communities have assessment mechanisms and demonstrate competence, honesty, and shared

21

understanding, which helps members decide who and what is trustworthy. Members can rely on each other's assessments of trust, and trust can be transferred to other members. For example, Jirotka et al. (2005) found that breast-cancer researchers rely on the knowledge of colleagues' performance in reading mammograms to assess the trustworthiness of sources, in addition to other elements of judgment. Epistemic communities are even more helpful when it is difficult to judge unknown data producers' skill based on the data. In this case, firsthand knowledge of the skills or values of other researchers affects the assessment of trust and reuse decisions (Zimmerman, 2008).

1.2.3. Data producers

In addition to seeing data producers as parts of CoPs, other information related to data producers can be used to judge trustworthiness and make decisions on reuse. Zimmerman (2008) found that when ecologists use other insights from the field to assess the trustworthiness of data sources, their first focus is on who originally collected the data. Competence, commitment, and the reputations of specific data producers are important assessment criteria (Van House, 2002; Van House et al., 1998). However, when ecologists used personal knowledge about data producers, the insights did not give an automatic acceptance of data and instead played a secondary role in data-reuse decisions (Zimmerman, 2008). Just as Faniel and Jacobsen's study (2010) proposed other factors influencing data-reuse decisions, the decision-and-assessment process is not simple, and one factor cannot solely provide total trust on data and acceptance. However, information about data producers helps lessen reusers' concerns about data quality (Zimmerman, 2008).

1.2.4. Individual knowledge, skills, and experiences

Reusers' personal knowledge, skills, and experiences also play a role in trust assessments. In many cases, reusers deal with uncertainty due to the challenges of data reuse described in the section above (e.g., missed contextual information and insufficient documentation). Knowledge and skills from reusers' own data-collection experiences help them not only understand the data, but also judge the data quality and trustworthiness (Borgman, 2007; Zimmerman 2008). Reusers' knowledge and understanding of the errors that can occur in data collection were key aspects of their ability to judge data quality in Zimmerman's study of ecologists, and judgments on the competence and commitment of data producers can also be based on reusers' perceptions and personal knowledge (Zimmerman, 2008).

1.2.5. Data elements

Trust also stems from factors in the data themselves, such as collection methods, measurements, or variables. Before reusers trust data, they need to know what is being observed and how. For instance, habitat biologists asked how data-collection instruments were chosen and how data producers calibrated the instruments before reusing them (Wallis et al., 2007). Zimmerman (2008) remarked that what is being observed is sometimes the primary source of trust, which is justified by how it is collected. Knowing how easy (or difficult) data are to collect affects reuse decisions. For example, if reusers in ecology were suspicious of a certain variable, they could choose to exclude it (Zimmerman, 2008). Faniel and Jacobsen (2010) also found that earthquake data reusers' understanding of how data producers collected and measured data increased their trust in data reliability.

Validity is another critical element related to trust judgments. Data reusers know that

problems might exist when data are collected. Knowing how problems were resolved during the collection (or experiment) processes helps reusers know what is valid and what is invalid, which increases their trust that the data was properly processed (Faniel & Jacobsen, 2010).

1.2.6. Information availability

Because assessing trust of data inevitably requires an in-depth understanding of their context (Jirotka et al., 2005), the amount of information that can be delivered by any means is important. As already discussed, information can be obtained through reusers' previous knowledge; their familiarity of artifacts and process; perceived competence or honesty of data producers; or direct interaction with colleagues, experts, or data producers (e.g., Birnholtz & Bietz, 2003; Van House, 2002). It can also be delivered through documentation, which is the most ideal situation because information contained within documentation is stable; yet the memories of data producers can be changeable, partially lost or distorted, while documentation can provide necessary information without human interaction. Faniel and Jacobsen (2010) found that earthquake engineering data reusers developed their trust by reviewing documentation.

1.2.7. Data repository

Previous studies found that reusers' judgment of trust is closely related to data repositories. Not all data come from repositories, and reusers still acquire data directly from their peers' laboratories or their colleagues. Nevertheless, when reusers get data from repositories, it is important to know how the data are processed (Carlson & Anderson, 2007; Yoon, 2014). In Carlson and Anderson's ethnographic studies (2007) on actual data-sharing practices, reusers wanted to know how the data was "cooked," and they did not trust them if they could not find out. How and to what extent to repositories can provide data transformation is a challenge not

only in eScience but also in digital curation communities (e.g., Chen, 2005; Groth, 2005) that are still being discussed and developed.

Data repositories in Carlson and Anderson's studies (2007) attempted to be explicit about these data processes. They also tried to address the issue of trust by recording and providing provenance information to reusers. Other organizational attributes, such as the integrity of repositories, transparency, reputation, and structural assurance that guarantees of preservation and sustainability were also identified as important trust factors in previous studies (Yakel et al, 2013; Yoon, 2014). One interesting aspect of reusers' trust in data repositories is the reusers' perceptions of the roles of repositories. Yoon's (2014) study suggested that users' awareness of repositories' roles or functions was one factor for developing users' trust. Sometimes what reusers assumed about the repository functions was incorrect, but what reusers believed still influenced their trust.

While trust in repositories has been discussed significantly through the efforts to build trusted digital repositories (e.g. *Trustworthy Repositories Audit & Certification: Criteria and Checklist* (TRAC), nestor: *Catalogue of criteria for trusted digital repositories*, Data Seal of Approval (DSA), and so on), whether end users will accept a repository with a solid record as "trusted" is an area that remains unanswered. Because "trusted digital repositories can be classified as 'trusted' primarily because they meet or exceed the expectations and needs of the user communities for which they are designed" (Prieto, 2009, p. 603), it is significant to prove repositories themselves as "trusted" by user communities.

These previous studies have investigated how reusers select data and how trust plays an important role in their decision-making process, but some have reported inconsistent findings.

For example, in the case study of the NeuroAnatomical Cell Repository (NACR), the identity of the individual who produced or authored the data was not as important as particulars about the data itself (Cragin & Shankar, 2006). Cragin and Shankar (2006) observed that understanding the roles of data and metadata characteristics would be more important in terms of affecting trust and reuse. In addition, while Faniel and Jacobsen (2010) reported that reusers developed their trust from documentation and not from other elements (e.g., interaction with colleagues, information about data producers, etc.), other studies found reusers' trust is associated with these factors. Disciplinary differences or types of data might cause these inconsistencies, and more research in different disciplines and across disciplines is needed.

2. Data curation and trust

As already noted, proper data curation is an integral part of data reuse, as it enhances data reusability. Over the past decades, much research has been conducted regarding long-term digital preservation and curation, most of which applies to the needs of cultural institutions, such as libraries, museums, and archives. More discussions directly related to the preservation and curation of data are now emerging, although as Duerr et al. (2004) argued, data stewardship (or curation) is still a fairly new concept.

2.1. Meaning of data curation

The term *digital curation* is increasingly being used with related terms such as *digital preservation* or *digital archiving*. A challenge in discussing this subject is the lack of a standardized language. As of yet, there is no international authoritative source of community-approved definitions of these terms (ICPSR, 2009). In addition, *curation* often means different things to different audiences because the term is embedded deeply in local usages (Beagrie,

2008; Shorish, 2012). A major barrier, therefore, is the different interpretations and usages of terminology by different individuals and disciplines (Beagrie, 2008).

Because of their expertise and longstanding tradition of preserving resources, archival communities have been called to add valuable perspectives on data curation issues (Akmon et al., 2011), and recently, some archival researchers have begun to pay more attention to data curation (e.g., Shankar, 2007). Library and Information Science (LIS) research has also been responding to the new service demands associated with data curation; and the term *data curation* started appearing regularly in LIS and archival science literature in the 2000s and has increased notably since 2004 (Weber, Palmer, & Chao, 2012). A majority of the literature has focused on what curation means in the context of a specific discipline and why data curation is important. Table 2 presents several examples of proposed definitions of data curation.

Table 2. *Definitions of data curation*

Source	Definition of Data Curation
DCC	Maintaining and adding value to a trusted body of digital information for current and future use
ICPSR (2009)	Value-added activities to make specific data as understandable and usable as possible to their community
JISC (2003)	All the processes needed for good data creation and management, and the capacity to add value to generate new sources of information and knowledge
Lord & MacDonald (2003); Lord, MacDonald, Lyon & Giaretta (2004)	The activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and reuse. (…) Higher levels of curation will also involve maintaining links with annotation and other published materials.
MacDonald & Lord (2003): from the	An active management of information, involving planning; the actions involved in caring for digital data beyond its original

workshop discussion	use; add value to data; managing, improving, enhancing data
Patel & Ball (2008)	The active management of digital data and research results over their entire scholarly and scientific lifetime, both for current and future use
Shreeves & Cragin (2008)	The active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education, which includes appraisal and selection, representation, and organization of these data for access and use over time

Several core concepts and agreements on the meaning of data curation emerge from a review of the definitions. First, data curation is a continuum of activities that support both current and future data use. Basic curation activities involve organizing, providing access, and preserving, while continuum activities may reach back from the creation of data to the reuse of preserved data. Second, in relation to the continuum of activities, curation is the active management of digital information. This active management (or care) of data over their lifecycle is the key to the reproducibility, reusability, and potential and long-term value of data (Rusbridge et al., 2005). Active management often requires the involvement of not only curators but also data producers (or creators). Curating research data does not simply mean curating sources but also curating "knowledge," and the involvement of the interaction between producers, curators, and consumers (or users) of data is a significant part of this process (Angevaare, 2009). Third, most of the definitions underscore "adding value" as an important aspect of curation. Curation functions add value to data so the data become easier to discover, more accessible, richer in content, and easier to reuse (RIN, 2008). This view of curation goes beyond the enhanced present-day idea of reuse; it enables the verification of scientific discovery and provides data platforms for future research (Rusbridge et al., 2005).

Active management of data is one method to add value to data. The Research Information

28

Network (RIN) (2008) argued that adding value constitutes more than providing a brief annotation about context, which includes the followings:

- Annotating data by adding descriptions or other contextual information

- Adding additional data (e.g., combining data from other sources)

- Aggregating and linking to other types of data to produce a new corpus or test bed for analysis

- Providing metadata to make it easier to discover, access, use, and curate research datasets

- Providing tools for manipulating and using the data

- Curating and preserving datasets.

These activities are the responsibilities of curators or data repositories, but they also require the engagement of other parties. The question of who should carry out those activities and when they should take place are central to data curation.

2.2. Data curation lifecycle and actors in curation activities

In practice, different domains might curate data differently, but every domain should preserve data using common practices, at least at higher levels. Previous studies have identified current data curation practices in different fields, such as ecological data (e.g., Beagrie, Chruszcz, & Lavoie, 2008; Borgman, Wallis, & Enyedy, 2006), geospatial data (e.g., Digital Preservation Coalition, 2009; Erwin, Sweetkind-Singer, & Larsgaard, 2009), social science data (e.g, Gutmann, Schurer, Donakowski, & Beedham, 2004), and CENS (Center for Embedded Networked Sensing) data (Wallis, Borgman, Mayernik, & Pepe, 2008), and they have discussed issues to be researched further. Data curation practices, however, are not nearly as generalized as

preservation practices are. Borgman et al. (2007) found the reason in the lack of a common integrated data infrastructure in many scientific fields, which often results in non-standardized local data-management practices. Akmon et al. (2011) argued that archival researchers can bring valuable perspectives to data curation because of their experience with selecting data, preserving context in order to maintain meaning, and recognizing that digital materials must be addressed early in the lifecycle.

Despite this support, there are several useful models of data curation or the curation lifecycle, and all of them aim to provide effective guidance, either to broad disciplines or specific domains, such as the curation lifecycles proposed by ICPSR (2002) and Green and Gutmann (2007), which illustrated key considerations at each step from data creation to sharing and long-term management, and e)Research lifecycle view of data curation proposed by Lyon (2007), which included formulating ideas and hypotheses for new knowledge extraction. The DCC curation lifecycle model is the most well-known model and is often adopted in practices to identify the necessary activities for data curation (e.g., Heidorn, 2011). The DCC model provides a high-level overview of the required stages for the successful curation of data from initial receipt through the iterative curation cycle (Higgins, 2008). This lifecycle consists of different layers of actions. A full lifecycle is comprised of "sequential actions" beginning with "conceptualize" and moving clockwise to the next step; "Occasional actions" are the actions outside the sequential actions and re-order the sequential actions when decisions are made to do so (Figure 1). These actions are activities that take place around digital objects (data) and can occur at any time during the curation lifecycle. The actions include: 1) preservation planning; 2) community watching and participation, such as maintaining a watch on appropriate community activities and participating in the development of standards, tools, and software; and 3) curation

and preservation, such as undertaking the management and administrative actions planned to promote curation and preservation.
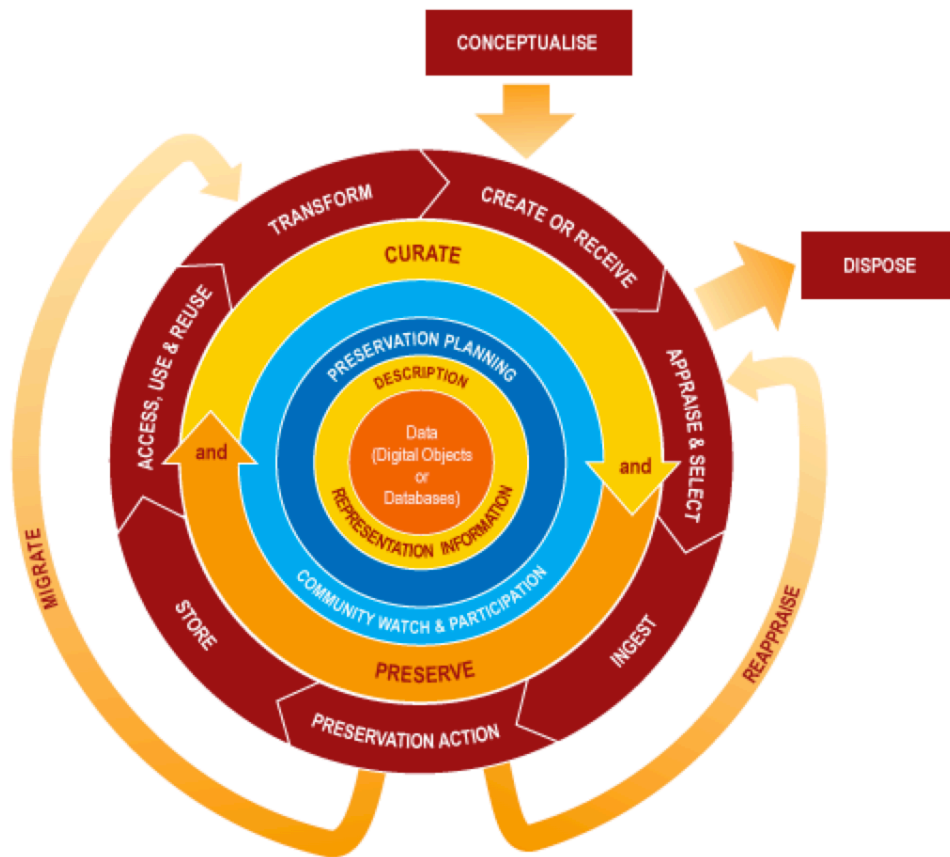
Figure 1. *DCC curation lifecycle model*

Although the DCC curation lifecycle model provides useful and generalized guidance, it is not always clear which stakeholders or actors must be involved in the curation lifecycle to implement all actions identified in the model, as the model focuses on *actions* rather than *actors*. For instance, based upon my analysis of the definition, I argue that data curation means active management, of which the involvement of data producers, curators, and reusers is the significant part. However, the role of data producers is invisible in this model, and the process of data

creation is encapsulated by "Conceptualise," when all actions of planning, collecting, processing, and cleaning data are (and should be) performed by data producers. A full understanding of data curation, thus, requires an understanding not only of actions but also of actors and their roles and responsibilities.

Previous researchers have identified different actors who contribute to curation in both micro and macrocosmic ways. Table 3 presents the summaries of the roles and responsibilities of the actors in data curation, including data producers (or creators), curators, reusers, institutions, funders, and publishers.

Table 3. *Roles and responsibilities of parties involved in data curation*

Roles	Responsibilities	Source
Producers or creators	- Manage data for the life of the project - Meet community standards for best practices and metadata - Data documentation: adequately describe the context and quality of the data and help others find and use the data; create codebooks and later interpret these codebooks; ensure all ethical and procedural documentation is in order - Data process: verify the accuracy and integrity of data sets - Develop and continuously refine a data management plan that describes the intended duration and migration path of the data - Comply with funder/institutional data policies - Protect the confidentiality of participants using proper documentation throughout the research	Humphrey et al. (2000) Lyon (2007) National Science Board (2005)

	process	
	• Meet appropriate legal standards for protecting security, privacy, and intellectual property rights	
Institutions	• Develop internal data management policies • Manage data on a short-term basis • Provide training and advice to support researchers	Lyon (2007)
Data archivists, mangers, curators/centers/repositories	• Establish actual data-archiving practices; participate in the development of community standards, including format, content (including metadata), and quality assessment and control • Manage data for long-term use; provide for the integrity, reliability, and preservation of data by developing and implementing plans for backup, migration, maintenance, and all aspects of change control • Meet standards for best practices • Protect rights of data contributors • Provide mechanisms for limiting access to protect property rights, confidentiality, and privacy; and enable other restrictions as necessary or appropriate • Provide tools for the reuse of data • Provide appropriate contextual information, including cross-references to other data sources • Promote repositories' services • Provide effective communication for the served community	Humphrey et al. (2000) Lyon (2007) National Science Board (2005)
Consumers/reusers	• Abide by license conditions/restrictions on use	Lyon (2007) National Science

	• Acknowledge data producers and curators	Board (2005)
	• Report significant errors to data managers or authors as appropriate	
	• Manage derived data effectively	
	• Reach a consensus on data center needs/structure for their user community and evaluate the quality of available centers	
Funding agencies	• Consider wider public policy and stakeholders' needs and develop a policy with stakeholders	Lyon (2007) National Science Board (2005)
	• Participate in strategy and policy coordination	
	• Monitor and enforce data policies	
	• Act as an advocate for data curation and fund expert advisory services	
	• Support the workforce capacity development of data curators	
	• Support interactions within and between communities to allow the development of robust community standards for digital data and interoperability and facilitate the development of community norms, customs, and expectations for digital research	
Publishers	• Engage stakeholders in the development of publication standards	Lyon (2007)
	• Link to data that support publication standards	
	• Monitor and enforce public standards	

While all of the actors of curation directly or indirectly influence reusers' experience of data, some actors can immediately impact users' experience with and trust of data due to their

roles, which are data producers and curators. Data producers are referred to as the scientists, educators, students, and others involved in the research that produces digital data (National Science Board, 2005), and OAIS (CCSDS, 2002; 2011) specifically defined a data producer as the entity who provides the information that will be preserved. In a data reuse context, a data producer can be defined as individuals or research teams that are engaged with data creation. Data producers play an important role in data lifecycles because they offer relevant contextual information about the data that can be lost when it is not documented at the time of creation. Although some information is inevitably lost in the process of knowledge transfer and documentation is inherently insufficient, engaging data producers in the process of knowledge transfer and documentation is one way to capture as much contextual information as possible for future interpretation. If the information is not received in a timely manner, data repositories will not be able to acquire, process, preserve, or disseminate data in an efficient and cost-effective way (Hedstrom & Niu, 2008). Martinez-Uribe and MacDonald (2009) argued for the need to engage researchers in curation activities, saying that producers' engagement provides valuable insight into the various stages of the data lifecycle and such activities help to gain the trust of producers, facilitating the process of data curation within data repositories at a point early on in the research lifecycle.

Another actor of curation who has direct impact on users is a data curator. Data curators perform value-adding activities to make data reusable, including cleaning, verifying, organizing, and documenting the data they have received. Because data producers may be insufficiently aware of their roles and responsibilities, cooperation between producers and curators is essential, and the process of depositing and documenting data is often a shared responsibility of data producers and managers/curators (National Science Board, 2005). Though data producers have

knowledge about data, data curators can provide expertise on how to document that information initially in an appropriate and effective way as well as identify the contextual information to document for future use. Data curators, thus, serve as intermediaries between data producers and reusers.

Data centers, archives, and repositories have tried to bridge producers and reusers by supporting knowledge transfer from producers to users through repositories. They have filled this intermediary support role using their expertise in managing and preserving data (Borgman, 2007; Green & Gutmann, 2007) and providing an infrastructure for data reuse through documentation and access services. The development of the Data Curation Profiles Toolkit (http://datacurationprofiles.org/) and the Data Documentation Initiative (DDI) (http://www.ddialliance.org/) are some examples of their work. Several studies (e.g., Martinez-Uribe & Macdonald, 2009) argued that building a trust relationship with both data producers and users is essential to perform the intermediary support role by engaging the two parties in data curation.

Because data reusers in a data curation lifecycle, who may include large scientific, education, and professional communities, are not passive recipients who just take data "as it comes" to them, it is important to note that their responsibilities include not only the conformance to any use rules based on use agreement, restrictions on use, or license conditions but also the evaluation of data and data curation activities (as well as repositories) and the reporting any errors. Reusers' assessments and input at the end of curation lifecycles can be triggers to improve current data curation practices and conceptualization stages in the lifecycle to plan activities so they meet users' standards. Due to the active roles and engagement of users in data curation, the actions and opinions of users can possibly impact other reusers of data.

While an examination of the activities performed by these actors confirms their impact on data reusers' experiences with the data, it also raises several questions that arise from the previous research and suggest what the various activities of data curation are. First is how to implement these curation activities for the actors engaged in the curation lifecycle. As discussed, the roles and responsibilities of data producers must grow in order to prepare "archive-ready data" (Hedstrom & Niu, 2008), which enable repositories to acquire data with the necessary descriptive information. Efforts have been made to help data producers with data management. A number of research libraries and data repositories provide data management plans for researchers. ICPSR,[1] MIT libraries,[2] Purdue University Library,[3] and the University of Virginia Library[4] are well-known institutions that provide guidance to researchers. There are also collaborative efforts, such as the Data Management Plan (DMP) Tool (https://dmptool.org/), which helps researchers to create "ready-to-use" data management plans.

Training and education in data curation are other important methods for supporting curation activities in the field through graduate-level instruction in data curation (e.g., Data Curation Education Program [DCEP] and data curation specialization within the Master of Science program at the University of Illinois, certificate programs or Masters-level courses at the University of Arizona, University of North Carolina at Chapel Hill, University of Michigan Ann Arbor, and Simmons University), and training programs for practitioners through workshops or

---

[1] Guidelines for Effective Data Management Plans, http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/

[2] Write a data management plan, http://libraries.mit.edu/data-management/plan/write/

[3] Data Management Plan (DMP), https://purr.purdue.edu/dmp

[4] Data Management, http://data.library.virginia.edu/data-management/

institutes (e.g., the Data Preservation Management Workshop [DPMP] hosted by MIT and DigCCurr Professional Institute from the University of North Carolina at Chapel Hill).

Another question is relevant to the need to understand how these curation activities performed by different actors influence users' perceptions and their trust in the data. Because users should also be very actively engaged in the data curation process rather than be passive receivers of information, listening to users' experiences and understanding their perceptions of data become even more important. Investigating the link between users' trust in the data and curation activities also helps to implement curation activities in a user-trusted way, which is a better way to serve users' needs. In the past, several studies have investigated reusers' trust in data in relation to the activities carried out by data curators, including the context of data repositories, as already discussed in an earlier section, where curation activities are traditionally performed, and other studies focused on a specific aspect of data to be trusted for curation, such as data integrity, quality, and provenance (e.g., Mayernik, Wallis, Pepe, & Borgman, 2008; Donaldson & Fear, 2011; Lemieux, 2014). Less research has been conducted regarding the curation of relevant activities performed by other actors and users' perceptions, which will require more empirical research with a holistic approach to understand the relationship between users' assessment on data, trust, and curation activities.

Chapter 3. Theoretical Framework

1. Definition of trust

The concept of trust has been widely studied in various disciplines, including sociology, social psychology, organizational behavior, marketing, and economics. Because researchers from different fields take varying approaches to understanding the concept of trust through their own disciplinary lenses and filters, a full consensus on the definition of trust has not yet been reached. Researchers have also argued about the difficulty of defining and measuring trust (Rousseau, Sitkin, Burt, & Camerer, 1998), since it is a vague term with an elusive definition (Gambetta, 1988). Meanwhile, these different conceptualizations of trust illustrate the multidimensional nature of trust.

Webster's dictionary defines *trust* as a) reliance on the character, ability, strength, or truth of someone or something; b) dependence on something future or contingent; and c) a property interest held by one person for the benefit of another (Trust, n.d.). As presented using the examples of definitions in Table 3, a number of studies have tried to define trust and proposed definitions of trust suitable to the context of their studies. One difference among various definitions of trust is trust as a belief vs. trust as a behavior. Psychologists working from personality theory have conceptualized trust as a psychological trait or state that individuals develop in varying degrees and thus have tended to see trust as a mental status or belief, whereas behavioral psychologists proposed a behavioral interpretation of the concept of "trust" by

equating trust with cooperation with others (Lewis & Weigert, 1985). Although these different perspectives would lead to fundamental differences in understanding trust and influence further development of a trust model, a few studies have not clearly indicated which perspective they followed and instead used mixed perspectives in conceptualizations. While the context of each study is different, and some definitions entail general trust whereas others tend towards specific trust, several phrases, such as confident expectation, positive expectation, or willingness to be vulnerable, were frequently used to describe the concept of trust.

Table 4. *Examples of trust definitions in previous literature*

Discipline	Source	Trust definition
Marketing	Anderson and Narus (1990)	A belief that another party (company) will perform actions that will result in positive outcomes for the party as well as not take unexpected actions that will result in negative outcomes.
	Anderson and Weitz (1989)	A belief that a trustor's needs will be fulfilled in the future by the actions undertaken by the other party
	Doney and Cannon (1997)	Perceived credibility and benevolence
	Magrath and Hardy (1989)	A belief that another person or thing (company) may be relied upon with confidence
	Morgan and Hunt (1994)	Confidence in an exchange partner's reliability and integrity
Organization -al studies	Jones and George (1998)	Experience of which is the outcome of the interaction of people's values, attitudes, and moods and emotions
	Mayer, Davis, and Schoorman (1995)	The willingness of a party 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
	McKnight and Chervany (1996)	Expectancies or beliefs. Expectancies reflect the future orientation of trust. Beliefs reflect the critical role that perceptions about the other party play in

trust

	Mishra (1996)	Willingness to be vulnerable based on the belief that the other party is competent, open, concerned, and reliable.
	Ring and Van de Ven (1992)	Confidence in the other's goodwill (adopted from Friedman (1991))
	Rousseau et al. (1998)	The intention to accept vulnerability based upon positive expectations of the intentions or behavior of another
	Sheppard and Sherman (1998)	The acceptance of the risks associated with the type and depth of the interdependence inherent in a given relationship
	Sitkin and Roth (1993)	A belief in a person's competence to perform a specific task under specific circumstances
	Williams (2001)	One's willingness to rely on another's actions in a situation involving the risk of opportunism
Sociology	Gambetta (1988)	A particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both before he/she can monitor such an action (or independent of his/her capacity ever to be able to monitor it) and in a context in which it affects his/her own action
	Zucker (1986)	A set of expectations shared by all those involved in an exchange; confidence or predictability in one's expectations
	Sztompka (1999)	A bet about the future contingent actions of others; a crucial strategy to deal with an uncertain, unpredictable and uncontrollable future; beliefs and commitment
Social psychology	Good (1988)	Based on an individual's theory as to how another person will perform on some future occasion, as a function of that target person's current and previous claims, either implicit or explicit, as to how they will behave
	Giffin (1967)	Reliance upon the characteristics of an object, the

		occurrence of an event, or the behaviors of a person in order to achieve a desired but uncertain objective in a risky situation
	Luhmann (1979)	Trusting behaviors imply allowing oneself to be in a potentially vulnerable position relative to another, while possessing some knowledge of the other that inspires trust in his/her goodwill, i.e., in his good intentions
	Rotter (1967)	An expectancy held by an individual or a group that the word, promise, or verbal or written statement of another individual or group can be relied upon
Economics	Baier (1986)	Accepted vulnerability to another's possible but not expected ill will (or lack of good will) towards oneself. It is reliance on the other's competence and willingness to look after, rather than harm, things one cares about which are entrusted to the other's care.
	Baker (1987); Lagerspetz (1992)	Beliefs that are not accepted on the basis of evidence and beliefs which in some cases might be highly resistant to evidence that runs counter to them
	Hertzberg (1988)	Attitude towards another person, without specifying whether he/she is trusted, as could be said that after judgment somebody is relied upon in certain aspects
	Lorenz (1988)	Trusting behavior consists of action that 1) increases one's vulnerability to another whose behavior is not under one's control, and 2) takes place in a situation where the penalty suffered if the trust is abused would lead one to regret the action
	Sako (1992)	An expectation held by one trading partner about another that the other behaves or responds in a predictable and mutually acceptable manner

T

he review of these trust definitions also reveals several common conceptualizations of trust, even though there is no universally accepted scholarly definition of trust across the disciplines. First, trust is associated with possible positive outcomes rather than negative outcomes, such as a

belief that another party will perform actions that will result in positive outcomes for the party (Anderson & Narus, 1990; Mayer et al., 1995). This positive expectation is linked to the risk taking: for example, 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), where risk in decisions is less pronounced, which brings greater optimism in the decision. The review also reveals that there is a relationship between predictability and trust, where trust is dependent upon predictable behaviors.

Mayer et al. (1995) proposed the definition, "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" (p. 712). This is one of the most frequently cited and influential definitions. It pertains to several common conceptualizations of trust proposed by others, such as positive expectations and vulnerability (in an association with risk). Later, in their study of a multidisciplinary view of trust, Rousseau et al. (1998) similarly reported that "confident expectations" and "willingness to be vulnerable" are critical components of all definitions of trust regardless of discipline and defined trust as "a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another" (p. 395). A similar definition again appears in Marsh and Dibben's (2003) work in their interdisciplinary reviews of literature on trust, saying that trust "concerns a positive expectation regarding the behavior of somebody or something in a situation that entails risk to the trusting party" (p. 470). Still, this identification of a widely shared meaning of trust does not imply that all operationalizations of trust reflect the same meaning. Nonetheless, the review of definitions makes it possible to understand the nature of trust on a broader level.

2. Pre-conditions of trust

Researchers have discussed the circumstances in which trust arises. Several conditions under which trust can be forged were described across disciplines, such as risk, uncertainty, vulnerability, and interdependence, though the words used to describe these conditions were not always the same. Defined as the perceived probability of loss, risk has been considered an essential component of the pre-conditions for trust (Deutsch, 1962; Lewis & Weigert, 1985; Rousseau et al., 1998; Rotter, 1967; Ring & Van de Ven, 1992; Sheppard & Sherman, 1998). Risk was also discussed in terms of the higher-level concept of uncertainty (Doney & Cannon, 1997; Gambatta, 1988; Lewicki & Bunker, 1996) because uncertainty can lead to various types of risk (Ring & Van de Ven, 1989). Uncertainty can also result from a lack of information (Giddens, 1990) or unknowable actions of another (Gambatta, 1988). Others have argued that trust is only relevant if a trustor is vulnerable to suffering a loss when betrayed (Doney & Cannon, 1997) and a trustor is willing to be in a vulnerable position (Blomqvist, 1997; Rousseau at el., 1998; Mayer et al., 1995). The final pre-condition of trust discussed by researchers is interdependence (or dependence). Interdependence (or dependence) means that a trustee holds the potential to satisfy a trustor's needs; thus, it occurs "where the interests of one party (trustor) cannot be achieved without reliance upon another (trustee)" (Rousseau et al., 1998, p. 395). Dependability and reliance (or reliability) has been emphasized by the large body of research as a necessary condition of trust (e.g., Blomqvist, 1997; Gambatta, 1988; Luhmann, 1979; Rempel et al., 1985; Rousseau et al., 1998; Rotter, 1971; Ring & Van de Ven, 1992; Weber et al., 2004). In their integrated model of trust, Kelton, Fleischmann, and Wallace (2008) confirmed these pre-conditions of trust—uncertainty, vulnerability (risk was discussed as a higher level of concept of

uncertainty and vulnerability) and dependence—and argued all of these conditions must be met for the questions of trust to become relevant.

3. Trust antecedent attributes

Trustworthiness is often used as a synonym for trust, but the definition is slightly different. Webster's Dictionary defines *trustworthy* as *worthy of confidence* and *dependable*. Whereas trust can be seen as a firm belief, trustworthiness emphasizes "being able to" believe or rely on someone or something. Seen as the "perceived likelihood that a particular trustee will uphold one's trust" (Kelton et al., 2008), trust can be directly influenced by the perceived trustworthiness of a trustee with other external influential factors, and thus trustworthiness is an important and useful concept to understand trust.

The first trust attribute commonly discussed among researchers is ability (e.g., Deutsch, 1960; Mayer et al., 1995; Sitkin & Roth, 1993). Ability refers to the skills, competence, and characteristics of a trustee that are influential in a specific domain. Others also discussed aspects of ability using different constructs. Competence—defined as the knowledge, expertise, or skills possessed by a trustee in order to fulfill the needs of the trustor (e.g., Butler, 1991; Kelton et al., 2008; Sheppard & Sherman, 1998)—or expertise (e.g., Giffin, 1967) are other constructs used. Although ability, competence, and expertise were used as synonyms, Mayer et al. (1995) proposed the use of the term *ability*, since *ability* highlights the task and situation-specific nature of the construct, while *competence* or *expertise* connote skills applicable to a single, fixed domain. Ability can be important to cognitive-based trust (see section 4.1.2. of this chapter for the definition), as it can provide a rational judgment on the trustee's capability to meet the needs of the trustor (Kelton et al., 2008).

The second identified characteristic is benevolence. Benevolence refers to the belief of a trustor that a trustee wants to do good work for a trustor (Doney & Cannon, 1997; Mayer et al., 1995; Sheppard & Sherman, 1998). Since this attribute represents positive intentions or feelings of a trustee towards a trustor, it is more related to affect-based trust (see section 4.1.3. of this chapter for the definition). Others used the term (positive) intentions or motivations (e.g., Deutsch, 1960; Giffin, 1967; Lewicki, McAllister, & Bies, 1998), which has a wider implication than personal orientation towards a trustee.

Integrity is another attribute of trust antecedents. Integrity is a trustor's perception that a trustee will adhere to principles acceptable to the trustor, and consistency of the trustee's past actions or a trustor's belief that a trustee has a strong sense of justice can affect a trustor's judgment of a trustee's integrity (Mayer et al., 1995; Sheppard & Sherman, 1998). Kelton et al. (2008) called this attribute *ethics*, which refers to the moral principles to which the trustee adheres. The term *ethics* is used as an umbrella term that includes not only integrity but also other ethical qualities, such as moral order discussed by Barber (1983), honesty by Muir (1994), or fairness by Butler (1991).

These three—ability, benevolence, and integrity—are the most commonly discussed attributes in previous literature. Known to be universally relevant, according to Mayer et al. (1995), many researchers have also adopted them as a starting point to develop their own framework in different contexts of trust research. Some researchers have augmented this set with additional attributes. Kelton et al. (2008) added *predictability* to these three attributes. Predictability means the degree to which the trustee's behavior conforms to expectations (e.g., Butler, 1991; Kelton et al., 2008; Sheppard & Sherman, 1998), and reliability (Gidden, 1990) or consistency (Butler, 1991) is often discussed as a synonym of predictability. Predictability can be

based on a trustor's observation of the past behaviors of a trustee or expectations of particular

social roles or functions of a trustee. Other attributes include identification and transparency,

proposed by Pirson and Malhotra (2011). Pirson and Malhotra (2011) slightly modified the

framework of three core attributes (ability, benevolence, and integrity) in the context of

stakeholders' trust in organizations by adding two more attributes: identification and

transparency. Identification refers to stakeholders' understanding of an organization's intention

or interests based on shared values and commitment (Lewicki & Bunker, 1996); transparency

refers to perceived willingness to share trust-related information with stakeholders. Though

transparency did not appear to predict trust in the results of their study, it is worth noting that

several scholars (e.g. Mishra, 1996; Tschannen-Moran, 2001) have argued for including

transparency as an attribute of trustworthiness.

In addition to trust antecedent attributes, or the attributes of perceived trustworthiness,

other attributes that influence the development of trust are identified. Depending on the research

context, different attributes are discussed, and it seems that there are no definite attributes across

disciplines or situations directly relevant to trust itself. Rather, these attributes generally are

related to the dimensions of trust. For example, the trustor's propensity to trust affects trust

development (Mayer et al., 1995). This attribute is related to personal trust, which views trust

from a psychological perspective. One who has higher propensity to trust in general would be

likely to trust more in a particular situation. McKnight et al. (1998) named this attribute *faith* in

humanity, in which one believes that others typically are well-intentioned and reliable. Other

attributes discussed in previous literature often relate to institutional or societal trust. These

attributes are more related to reputation; as Kelton et al. (2008) explained, a trustor is more likely

to trust if others also trust the trustee.

4. Types and levels of trust

Researchers have identified multiple types and levels of trust using different categorizations or criteria. These identified types and levels of trust have been used by a number of trust researchers as a basic framework to analyze certain approaches to trust. In this section, I categorize different dimensions and levels of trust from the previous literature, although sometimes the same types of dimensions or levels were described using different names or meanings. Each types and levels of trust is not necessarily mutually exclusive, and each can complement another to increase trust in a given situation.

4.1. Types of trust

4.1.1. Calculus-based or situational trust

Influenced by marketing and economics, researchers using calculus-based trust (Rousseau et al., 1998) or situational trust (Blomqvist, 1997; Sitkin & Roth, 1993) explain that trust can emerge based on a trustee's rational choice when a trustor perceives that the trustee will perform beneficial actions. Borrowing from Barber's (1983) argument, Rousseau et al. (1998) explained that this type of trust can be derived from credible information regarding the intention of another, which may be provided by reputation or certification. Blomqvist (1997) called it situational trust because it is likely to happen when a trustee behaves in an expected way because of factors external to the actor: for instance, when there is no viable choice and a trustee is thus forced to behave in an expected way.

4.1.2. Cognitive trust

Cognitive trust (Blomqvist, 1997) or cognition-based trust (McAllister, 1995; McKnight,

Cummings, & Chervany, 1998) may have some overlap with calculus-based trust, since it is often explained that trust can originate from "good rational reasons" (Lewis & Weigert, 1985, p. 972), for instance, why trusting a trustee (either a person or organization) results in merit. However, rather than having situational cues, it is grounded in a trustor's cognitive judgments of a trustee's competence or reliability (McAllister, 1995) or other cognitive cues such as a first impression (McKnight et al., 1998).

4.1.3. Emotional or affect-based trust

Emotional trust (Blomqvist, 1997) or affect-based trust (McAllister, 1995) is known to be founded in affective bonds among individuals. Lewis and Weigert (1985) also said trust (or trusting behavior) can be motivated by "strong positive affect for the object of trust" (p. 972). Affective and cognitive trust are not mutually exclusive, and it is possible for them to exist at the same time. According to Lewis and Weigert (1985), trust in everyday life is a mix of emotions and rational thinking.

4.2. Levels of trust

4.2.1. Personal or individual trust

Personal trust (Blomqvist, 1997)—in other words, individual trust (Kelton et al., 2008; Kini, 1998; Sitkin & Roth, 1993), personality-based trust (McKnight et al., 1998; Worchel, 1979), or dispositional trust (McKnight & Chervany, 1996)—focuses on individuals' personalities as a determinant of trust and sees trust as a purely psychological trait. Lewicki and Bunker (1996) referred this level of trust as "a belief, expectancy or feeling that is deeply rooted in the personality and its origins in the individual's early psychological development" (p. 115).

Bowlby (1982) and Erikson (1964) explained that trust develops during childhood (e.g., as an infant seeks and receives help from benevolent caregivers), which leads to a generalized expectancy or tendency to trust others (Rotter, 1971). However, Mayer et al. (1995) argued that this psychological attribute is not actually trust and used the term "propensity to trust," which influences whether an individual will extend trust in a particular instance.

4.2.2. Interpersonal trust

Interpersonal trust (Kelton et al., 2008; McKnight & Chervany, 1996; Sitkin & Roth, 1993; Worchel, 1979) or relationship trust (Kini, 1998) is the most common approach to trust, and it has been used by a number of researchers as a framework for their research. Researchers in psychology and sociology often see trust as an interpersonal relationship or social tie between a specific trustor and trustee. Mayer et al. (1995) defined trust as a relationship between a trusting party (trustor) and a party to be trusted (trustee), and McKnight and Chervany (1996) also referred it as two or more people (or groups) trusting each other in a specific situation. Rotter (1971) saw interpersonal trust as "an expectancy held by an individual or a group that the word, promise, verbal, or written statement of another individual or group can be relied on" (Rotter, 1971, p. 444). At this level of trust, the trustor usually has several attitudes toward the trustee, such as an expectation of the trustee's competence (Blomqvist, 1997), a sense of the risk associated with assuming or acting on such expectations (Lewicki et al., 1995), and a willingness to increase his or her vulnerability to the trustee (Zand, 1972; Mayer et al., 1995).

4.2.3. Relational trust

Relational trust (Kelton et al., 2008; Rousseau et al., 1998) can be seen as a form of interpersonal trust because it also focuses on the relationship between a trustor and trustee. The

trust-flow mechanism of relational trust, however, differs from that of interpersonal trust; in relational trust, the relationship is the moderator of trust. Unlike the approach of interpersonal trust—trust as an attitude or behavior from one to another—the approach of relationship trust sees trust as an emergent property of the relationship as a whole (Kelton et al., 2008). Rather than conceptualizing trust as expectancy or confidence, trust is an ongoing social practice; instead of trust flowing from a trustor to a trustee, it emerges and is derived from their repeated interaction over time (Rousseau et al., 1998). Thus, reliability and dependability in previous interactions can create and increase a trustor's positive expectations or beliefs about a trustee's intention (Rousseau et al., 1998; Lewicki & Bunker, 1996). Affect-based trust can be relevant to relational trust. McAllister (1995) explained that emotion can play a role in this repeated interaction as frequent, long-term interaction can lead to the formation of attachments based upon reciprocated interpersonal care and concern.

4.2.4. Institutional trust

Institutional trust and societal trust were defined somewhat similarly in previous literatures, and the terms *institutional trust* (or organizational trust) and *societal trust* were used in a mixed way. This literature review differentiates institutional trust and societal trust, placing societal trust as a broader dimension and not specific to one institution, following Kelton et al.'s (2008) definition of societal trust.

From the approach of sociology and economics, institutional trust focuses on the development of trust between individuals and institutions. Institution-based trust (McKnight et al., 1998; Rousseau et al., 1998; Sitkin & Roth, 1993) usually refers to the trust founded upon institutional structures or properties in the situation, not on a person. McKnight and Chervany

(1996) argued that this type of trust is not founded on a person associated with the institution, but others disagreed by saying that trust in an organization can stem from a person associated with the trusted parties, e.g., the owner's personality (e.g., Zucker, 1986). A specific organizational culture (e.g., a strongly centralized decision structure) or an organization's good reputation or resources can also create a feeling of trust, but this is not the same feeling as trust in a specific person (Halinen, 1994; Zucker, 1986). Trust can also be placed in functions rather than people (Luhmann, 1979). Blomqvist (1997) named this "system trust" and explained that system trust arises when a system is assumed to be functioning in a predictable way (e.g., the way bureaucratic sanctions are expected to function) (Luhmann, 1979) or when a trustor feels security about situations because of structural assurance, such as guarantees, regulations, or the legal system (e.g., a contract or promise, formal certification of expertise, etc.) (Rousseau at al., 1998; Sitkin & Roth, 1993). Institutional trust or system trust is known to serve as a substitute for interpersonal trust when the roots of interpersonal trust are not available: for instance, in the case of an absence of direct personal experience (Blomqvist, 1997; Sitkin & Roth, 1993).

4.2.5. Societal trust

Societal trust can be applied more broadly to a society. Kelton et al. (2008) and Worchel (1979) emphasized the importance of societal trust as a proper functioning of a society. Luhmann (1979) highlighted the role of societal trust in enabling people to cope with society. System trust also matters at a societal level, and it is based on the conformation to expectations of socially accepted behavior and social control, e.g., one's expectations of another (Kelton et al., 2008). Thus, similar to institutional trust, societal trust helps to ensure the proper functioning of society by encouraging individuals to place their personal trust in unknown parties, with which they might have low familiarity or few interactions. Kelton et al. (2008) also suggested that research

in this area focused on cultural features and social institutions that contribute to system trust, such as family structure or legal systems and the resulting benefits to society.

5. Trust development

Although early research on trust considered trust as static (e.g., economics) and social psychologists often saw trust with an all-or-nothing view, as if trustees either completely trust or distrust, a large number of 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 et al., 2008; Rousseau et al., 1998) who see trust as dynamic have described the process whereby trust is developed. Three phases of trust—*building* (where trust is formed), *stability* (where trust already exists), and *dissolution* (where trust declines)— were explained by Rousseau et al. (1998), but the three phases of trust were limited to providing the simple status of trust existence.

A more developed version of the trust-building process was proposed by Doney and Cannon (1997) and Kelton et al. (2008). Doney and Cannon (1997) presented a four-step trust-building process consisting of *prediction*, *capability*, *intentionality*, and *transference*: similarly, Kelton et al. (2008) provided 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 the trustee's credibility and benevolence, or it can also be based on the trustee's past behavior and promises. The *capability* process involves determining the trustee's ability to meet obligations, while the *intentionality* process involves interpretation and assessment of the trustee's motives, out of which trust emerges. Kelton et al. (2008) used the stage of prediction more broadly, including Doney and Cannon's (1997) prediction, capability,

and intentionality stages, and explained that prediction is based on the past behavior of trustees. The attribution process refers to the assessment of the underlying quality or motivations of trustees based on observations, just as the intentionality process does. The *bonding* stage proposed by Kelton et al. (2008) is the emotional development of a trustor–trustee relationship, which strengthens the relationship. The next stage is called *reputation* (Kelton et al., 2008) or *transference* (Doney & Cannon, 1997). This stage refers to the awarding of trust on the recommendation of others, and in this stage, trust is further developed and transferred to other parties as a "proof source." Lastly, Kelton et al. (2008) added *identification* as a final stage, which is developed when trustors and trustees share a common identity, goals, and values.

While these earlier studies addressed the development of trust through the accumulated experience of various users, McKnight et al. (2002) focused specifically on the formation of initial trust, which was developed in the context of e-commerce. Initial trust happens before the stage of prediction in the process of trust development and is formed when trustors do not have credible information, knowledge, or emotional bonds, nor have they had any direct interaction with a certain system or environment. McKnight et al. (2002) state that credible information is only gained after a trustor observes the trustee's trustworthiness and related behavior; during this period of time, the trustor develops trust (or distrust) of the trustee (a system or vendor). While it is significant to understand the trust-building process with accumulated user experience, understanding initial trust should not be neglected as users might often run into unfamiliar trustees, services, or other parties.

6. Trust in digital information

The concept of trust has recently been applied to information in the context of information exchange and use. Hertzum, Andersen, Andersen, and Hansen (2002) argued that trust plays a key role whenever people exchange information. Similarly, Kelton et al. (2008) said trust in information is useful to provide adequate explanations of information use because traditional models of information quality, which focus on attributes of the information itself, does not fully explain information use.

Trust in information is often used alternatively or interchangeably with information credibility, and in some cases, credibility has been conceptualized as a component or type of trust (e.g., Giffin, 1967). However, Tseng and Fogg (1999) differentiated credibility from trust; credibility means believability, while trust indicates dependability. According to their differentiations, credibility can be seen as users' perceived information quality or users' assessment of information quality. Trust, on the other hand, is more related to the willingness to depend on the truthfulness of information. Still, credibility is closely related to trust. Hertzum et al. (2002) thus defined credibility as the perceived quality of a source or piece of information and argued that establishing information credibility is a matter of determining to what extent one is willing to place trust in it.

Substantial research has been conducted on the information credibility of websites (e.g., Chesney, 2006; Flanagin & Metzger, 2000; Fritch & Cromwell, 2001; McKnight & Kacmar, 2006; Rieh & Belkin, 1998), but few have used the specific term "trust," and there seems to be little agreement on the influential factors and other aspects of trust in information. Several researchers (e.g., Castelfranchi et al., 2003; Corritore et al., 2003; Lucassen & Schraagen, 2011;

Kelton et al., 2008) adopted a previously developed concept or model of trust to information and added other factors fit to the specific context of information.

For pre-conditions of trust, Kelton et al. (2008) modified previously defined pre-conditions of trust—risk (encapsulating uncertainty and vulnerability) and dependability—and the proposed lack of standards for ensuring the quality of information, the potential harm from using poor information, and users' need for the information to support decision making as pre-conditions of trust in information. Because of the large volume of information in the digital environment, the lack of standards can result in uncertainty of information quality. When users judge information quality, they are also aware of potential harm that can result from using poor information (vulnerability). Finally, users are also aware of their dependence; information possesses the potential to satisfy their needs supported by any facts, references, or personal knowledge in their decision-making process (Kelton et al. 2008).

Previously identified trust antecedent attributes, ability (or competence), benevolence, and integrity, are also modified in the context of information, which links to information quality. In information, ability can be understood as accuracy or correctness (Lucassen & Schraagen, 2011; Kelton et al., 2008), which means the information is free from error. Lucassen and Schraagen (2011) also added completeness of information in addition to accuracy. Benevolence is understood as objectivity (Kelton et al., 2008) or neutrality (Lucassen & Schraagen, 2011) of information, which means the information is free from deception or distortion. The ethical aspect, *integrity*, indicates validity in information, which refers to the use of accepted practice in creating information, such as using a sound method, verifiable data, or correct citation sources. In addition to these three attributes, Lucassen and Schraagen (2011) added the source of

authority, which relates to the creators or owners of information. Kelton et al. (2008) mentioned that the stability of information since the advent of digital information is fluid and can be altered.

As previous studies have identified, personal trust is one of the other influential attributes of trust in information (Corritore et al., 2003; Kelton et al., 2008). This personal tendency to trust, or generalized positive expectancies, applies to information, and affects whether users would receive information with suspicion or not. While this argument is rooted in the psychological stance of trust, Lucassen and Schraagen (2011) argued trust judgments depend on users' expertise, specifically domain expertise and information skills. Domain expertise is known to be an important element of information credibility (e.g., Adelson, 1984, Chesney, 2006), and it helps users to assess information accuracy, neutrality, or completeness. Information skills are more generic skills of processing information and do not require domain expertise. Situational trust and societal trust appear as relevance (Kelton et al., 2008) or scope (Lucassen & Schraagen, 2011) and recommendations (Kelton et al., 2008). Relevance is another dimension of information quality (Rieh & Belkin, 1998) that refers to the degree to which information meets the requirements of users' needs. In addition, recommendations are one of the evaluation sources of information that rely on others' reviews or assessments of information.

The review of previous literature on trust definitions, dimensions, and development reveals its possible applications to different contexts. As previous studies have suggested, trust plays an important role in data reuse when researchers assess data reusability, earlier trust research has great potential to contribute to investigations of the complex nature of trust in data reuse. Different types of trust can be associated in this context, because trust development can be influenced by sources (data), people (data producers or managers), institutions (repositories), or society (a research community or academia as a whole).

Chapter 4. Research Methods

As this study concerns individuals' perspectives and thoughts regarding their experience, a qualitative research method was appropriate. In general, qualitative research is concerned with meaning. Qualitative researchers are interested in how people experience and make sense of the world. They try to develop an understanding of people's experiences and actions as they live through situations, based on the individuals' own perspectives (Elliott, Fischer, & Rennie, 1999; Willig, 2008). Thus, qualitative methods are suitable for capturing people's perspectives, including their thoughts, attitudes, and emotions.

1. Epistemological Stance

Not all qualitative research is rooted in the same philosophical stance. Positivistic, phenomenological, hermeneutic, pragmatic, critical, and postmodernist traditions influence different epistemological positions among qualitative researchers (e.g., empiricists or social constructionists) (Elliott et al., 1999). This study specifically follows an interpretive approach as its methodological stance.

The foundational assumption of interpretive research is the knowledge we gain or the reality we know is socially constructed, and making meaning is a social process (Willis, 2007). As access to reality is possible only through social constructions (e.g., language, consciousness, and shared meanings), interpretative research rejects an objective or factual reality or situation. The phenomena that researchers understand are through subjective or inter-subjective meanings

that participants assign to their world or experiences (Orlikowski & Baroudi, 1991; Walsham, 1993). Interpretive research focuses on understanding values, meanings, beliefs, thoughts, and the characteristics of specific phenomena in a particular context from the study participants' viewpoint (Leininger, 1985). The goal of interpretive research is thus more of an understanding of a particular situation or context than the generalized meaning or the discovery of universal laws or rules (Schwartz-Shea, 2012; Willis, 2007).

As it studies real-life situations as they unfold, interpretive research is non-manipulative, unobtrusive, and non-controlling, and it does not set out to test key concepts defined before the research has begun. An interpretive researcher may study established relevant literature and develop a sense of how these concepts are discussed; following a naturalistic and holistic approach, they do not bring their own scientific definitions with them to the field settings in order to test the accuracy of those understandings. Rather, they try to understand how the concepts that are key to a particular setting are used in the field and let this understanding "emerge from the field" (Schwartz-Shea, 2012).

Although interpretive research focuses on more context-specific meanings rather than the generalization of a finding to a population, some findings can influence the questions asked in a new study on the same topic. Willis (2007) argued that researchers may find similarities between the studied settings and other settings, and Orlikowski and Baroudi (1991) reported that interpretative research can be used outside of the context in which it was conducted, as it can inform other settings.

In library and information science (LIS), interpretative research was introduced in the late 1980s, when a positivist approach was more dominant (Hansson, 2005). However, as many

research problems and practices in LIS can be described as interpretive, the interpretative approach has been widely used in information science research in the area of information management, information systems, information retrieval, library practices, and reference theory (e.g., Bendiktsson, 1989; Budd, 1995; Capurro, 2000; Cornelius, 1996; Radfor, 1992; Walsham, 1995). Some argued for the usefulness of the interpretative method. Budd (1995) said that interpretative research (hermeneutics), in particular, provides a more holistic view, and Cornelius (1996) argued that it offers a deep insight into practices: "[A]ll participants in a practice are inevitably theorizing about their place and developing or sharing interpretations of it that reflects what they see the practice as being" (pp. 207–208). Further, "The interpretive approach offers practitioners the means to recapture research and theory in the field and to harness it to improved practice and enhanced and broader understanding" (Cornelius, 1996, p. 215).

2. Research Design

I conducted in-depth, semi-structured interviews with researchers in two disciplines (social work and public health). I selected potential study participants through purposive sampling of major database searches in the domain of social science. I also adopted several strategies to analyze the interview data and help me to move from description to interpretation.

2.1. Study sample

Previous studies on data reuse and reusers' trust revealed some inconsistencies in their findings; sometimes it is not clear whether the inconsistency is due to the data types or disciplinary differences. In order to recruit as homogeneous a sample as possible and not to have mixed influential factors, this study limited the data type to quantitative social science data and tried to recruit researchers who have reused data.

As the aim of this study is to provide details about the perceptions and understandings of a particular group, this study employed purposive sampling with a small sample size and sought individuals who could provide a detailed account of their experience. Purposive sampling is one of the most important kinds of nonprobability sampling to identify relevant participants (Welman & Kruger, 1999). Popay, Rogers, and Williams (1998) argued that "randomness and representativeness are of less concern than relevance" in qualitative research and the real question is, "does the sample produce the type of knowledge necessary to understand the structure and processes within which the individuals or situations are located" (p. 346). Purposive sampling is also known as an appropriate sampling method in interpretive research (Polkinghorne, 1989; van Manen, 1997) because the depth of data comes from the richness of the participants' experience of the phenomena under investigation (Smith, 2004). I also adopted strategies for conducting purposive samplings and for selecting cases from the previous studies. For instance, Flick (2009) suggested identifying particularly typical cases; making choices according to the intensity of the interesting features, processes, and experiences; and selecting only those cases in which the relations to be studied are especially clear or important. Convenience, or ease of access under the given conditions, can also be an important criterion for selection, as it can reduce the effort and time required for research (Patton, 2002).

In order to identify individuals who have data reuse experience, I used data citation tracking from major databases. While 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). Tracking data citation is a challenging process and may have limitations, but it is still the most effective way to identify data reusers.

I used major databases for publications, which are provided in the UNC Library E-Research tools (http://eresources.lib.unc.edu/eid/) as a starting point. The databases include EBSCOHost, SAGE Journals, ProQuest Social Science, and ERIC. Data reusers were identified from a keyword search in the full text from each database, using the search terms "secondary data" or "secondary analysis"; *secondary data* or *secondary analysis* are more commonly used than *reuse* in the social sciences (Gleit & Graham, 1989; Hinds, Vogel, & Clarke-Steffen, 1997). These searches were limited to journal publications and conference proceedings published in the United States for the convenience of conducting interviews, following reverse chronological order (starting from the most recent ones). In case the authors' contact information was not provided in the publications, I conducted an additional Google search to collect the authors' contact information (e-mail addresses and office phone numbers). For articles that were written by multiple authors, either the information of the corresponding author or the first author was collected. If the first author did not directly work with data sets, I then asked the first author to identify the appropriate person for this research and contacted him or her.

Researchers in various disciplines were identified from the search, but this study focused on two disciplines: public health and social work. The two disciplines were chosen for this study for three reasons. First, from the search, both social work and public health presented the largest number of researchers among other disciplines with various quantitative data used. This is empirical evidence that both disciplines have data-reuse cultures while providing enough potential study participants that I could contact. In addition, the use of secondary data in research in both disciplines was discussed by scholars. Boslaugh (2007) argued that the role of secondary data analysis is increasingly important in public health research and practice, and Guest and Namey (2014) also said that secondary data analysis plays a key role in modern public health

research. Secondary data has received less attention in social work despite social work researchers' recognition of secondary analysis among many other methodologies. However, the use of secondary data has been growing in recent years as social work has become more active in federally funded research (Sales, Lichtenwalter, & Fevola, 2006). Finally, the disciplines share similar disciplinary characteristics, which were helpful in recruiting a homogeneous sample; both disciplines have a professional orientation in their research, and they use several of same data sets due to shared interests in certain research topics (though they have different approaches to the same research problems).

2.2. Unit of analysis

The unit of analysis for this study is an individual and his/her data reuse experiences. Participants were encouraged to draw from their past experiences to answer interview questions, which are not necessarily limited to one single case. Data reusers' trust can be developed based on their cumulative experiences, and there might be a number of other experiences relevant to their experience, even though the experience is not always directly related to a single incident.

2.3. Data collection

Publications

All of the publications that I used to identify data reusers were collected. The purpose of collecting the publications is to gather information about datasets that were used in each publication. I read these articles to understand the type of research that was conducted, to determine what kinds of data sets were used in their research, and to see what comments the authors made about the ease or difficulty in using the data before I conducted interviews. This

process was followed to facilitate communication with participants and enhance my understanding; no further analysis was conducted on these publications.

Narrative data

I collected narrative data through in-depth, semi-structured interviewing. Kvale (1996) described the qualitative interview as "literally an inter view, an interchange of views between two persons conversing about a theme of mutual interest," where the researcher attempts to "understand the world from the subjects' point of view, to unfold the meaning of peoples' experiences" (pp. 1–2). Semi-structured interviews are known as nondirective, as questions (usually open-ended) asked by researchers are used as triggers for further conversation (Willig, 2008). Semi-structured interviews allow researchers more freedom to probe interesting areas that arise, and they can follow participants' interests or concerns. This flexibility of coverage allows the interviews to go into novel areas, which helps to produce richer data (Smith & Osborn, 2008). Because of this flexibility, however, it is difficult but important to keep a balance between maintaining control of the interviews and investigating predefined topics. A carefully constructed interview agenda is thus critical to keeping this balance and not losing the original research questions (Willig, 2008).

The total number of potential participants identified from the database search and contacted for an interview was 229 (public health: 123; social work: 106). A total of 58 researchers responded to the e-mail invitation to the study, and the response rate was 25.3%. Among them, a total of 38 researchers who affirmatively responded were interviewed (37 total interview sessions were conducted as two researchers participated in one interview).

Interviews were scheduled at the convenience of the research participants from May to

September 2014. Given the diverse geographic distribution of data reusers, all interviews were phone interviews. The duration of interviews varied from 40 minutes to 95 minutes, but the average length was 60 minutes. If further clarifications or explorations emerged from the interview data, a follow-up interview was scheduled, depending on the participants' willingness to elaborate more on their experiences.

In order to encourage the interview process to stay as close to the participants' lived experience as possible, nondirective open questions were asked. Interview questions focused on the following (See Appendix B for the interview guide):

- Research participants' reuse experience, including the process of data discovery, initial selection, and trust criteria of data before reusing, any changes in their trust judgment during the actual reusing process, and any factors that influenced this process.
- Research participants' thoughts and perceptions of data, trust, and their experiences.

Table 5. *Summary of collected data*

Collected data	Description	Collection method
Narrative data (Primary)	Content of semi-structured interview from individual participants (data reusers)	One-to-one phone interviews. All were audio recorded and transcribed.
Publication (Supplementary)	Published journal or conference papers by participants	Downloaded from the major databases (e.g., EBSCOHost, SAGE Journals, ProQuest Social Science, and ERIC)

In order to encourage participants to speak freely and openly, I used different probes during the interviews. As Bernard (2000) argued, the key to successful interviewing is to use

probing effectively by stimulating participants to produce more information without injecting the researchers' own biases into the interactions. Semi-structured interviews allow the use of probes, which helps to maximize the potential for interactive opportunities between participants and researchers. Bernard (2000) described four types of probing: (a) "what" questions, which prompt participants while minimizing the influence of the interviewer; (b) silent probes, which involve interviewers remaining quiet and providing wait time for participants to continue; (c) echo probes, which consist of interviewers repeating participants' most recent comments and asking them to continue; and (d) "uh-huh" probes, in which interviewers provide affirmative reactions to participants' responses in order to encourage them to continue. In addition to encouraging participants, probing helps the interview process in many ways: It can evoke valuable and more complete information (Bailey, 1987), it allows for the clarification of issues raised by participants (Hutchinson & Skodal-Wilson, 1992), and it helps participants recall information in response to questions (Smith, 1996).

All interviews were audio-recorded and then fully transcribed. I transcribed one-third of the interviews, and a professional transcriber did the remaining two-thirds. In order to protect the study participants and their confidentiality, the names of the data sets and the names of relevant researchers mentioned during the interviews were kept anonymous in all transcripts.

2.4. Data analysis

Several strategies can be used to analyze interview data that allow a researcher to move from description to interpretation, capturing initial thoughts to generating themes, through iterative and inductive cycles (Smith, 2007). The current research involved the following steps:

Close reading and re-reading

This process involves immersing the researcher in the original data. In this stage, I first read the printed transcripts while listening to the recorded audio. This process gave me an opportunity to be more familiar with the transcripts as well as perform a final proofread of the transcripts. Then I re-read the transcripts and marked important phrases and words.

Initial noting and coding

This step allows for the examination of semantic content and language use on an exploratory level and produces comprehensive notes and comments (descriptive, linguistic, and conceptual) on the data. In this stage, I wrote descriptive or exploratory notes in the margins of the printed transcripts. I also came up with the labels that best describe the participants' experiences, thoughts, or feelings and coded each transcript using the labels. After I worked on the paper transcripts, I transferred the transcripts using a qualitative data analysis tool, NVivo for Mac, developed by QST International, to facilitate further analysis. NVivo is used often by academics and other organizations and is known as an effective tool for analysis.

Developing emerging themes

The researcher maps the interrelationships, connections, and patterns between exploratory notes and produces a concise main statement based on the comments. While these initial notes are loose and open-ended, emerging themes capture and reflect both participants' original words and thoughts and the researcher's interpretations. In this stage, I concentrated on the patterns across interviews through comparisons and contrasting and developed top labels to categorize codes used during the initial analysis. I renamed the top or sub-labels or regrouped the

codes during this process. I also exported Microsoft Word documents of each code, and read them again while highlighting good quotations and writing memos.

Searching for connections across identified emergent themes

This method involves the development of a process for the researcher to think about how themes fit together, and the researcher is encouraged to explore and innovate in order to organize and map the themes. In this final stage, I organized categorized themes based on my understanding and tried to fit them into the big picture of data reusers' trust research as well as data curation practices.

2.5. Ensuring trustworthiness of qualitative research

There have been a number of studies regarding ways to assess the quality and validity of qualitative research, determining what quality and validity mean in qualitative research, and identifying how to develop quality standards (e.g., principles for trustworthiness) (Lincoln & Guba, 1985; Morrow, 2005; Morse, Barrett, Mayan, Olson, & Spiers, 2002; Packer & Addison, 1989; Rolfe, 2006). Because qualitative research is fundamentally different from quantitative research, a number of researchers have argued that the terms *reliability* and *validity* are not as pertinent in qualitative research as they are in quantitative research (e.g., Altheide & Johnson, 1998; Leininger, 1994). A new way of discussing the quality of qualitative research was suggested by others, and Lincoln and Guba (1985) proposed the term *trustworthiness*, as a parallel concept to reliability and validity. It consists of credibility (confidence in the 'truth' of the findings), transferability (the findings have applicability in other contexts), dependability (the findings are consistent and could be repeated), and confirmability (a degree of neutraility or the extent to which the findings are shaped by the respondents and not by researchers' bias,

68

motivation, or interest).

In order to ensure and demonstrate the trustworthiness of the study, I conducted a member check during the data analysis. Member checking refers to the practice of sending or bringing written materials involving the people studied back to them, which are commonly transcripts of interviews or segments of research manuscripts. The primary intention of member checking is to see whether the researcher developed the points from the perspectives of participants who are native to the situations (Schwartz-Shea, 2012).

During the data analysis, I contacted two participants for member checking to verify the experiences that they detailed during the interviews. I shared both parts of the interview transcripts along with summaries of my understanding. We discussed the accuracy of the excerpts, my understanding, and additional thoughts or insights that were relevant to the experience. This process was essential to confirm my analysis of their experiences.

3. Terms frequently used in this study

Data. The term *data* is used as an overarching term referring to different notions and examples. Data include a conceptual notion in a generalized context (e.g., "social science data" or "importance of data in research") and in examples that participants used and mentioned (e.g. "I use xx data," "I download the data from…"). Data can consist of multiple files, including data files, documentation, and other associated information. Some participants also used the term *data set*, and this term remained in the direct quotations. Some participants also used a plural form, data sets, when they referred to multiple data, e.g., "combining data sets."

Original investigators. I use *original investigators* to refer to data producers (as

participants referred to them as the "PI" in an original study). Original investigators can be individual researchers (including research teams that consist of a group of individual researchers) and institutions that are responsible for designing and creating data. When I need to specifically refer to a type of original investigator, I use *individual researchers* or *institutions*.

Original study. *Original study* refers to the study that initially produced the data reused by the participants.

Chapter 5. Research Results

The results are organized in several parts. In section 1, I report the demographic

characteristics of the participants. Section 2 presents the underlying motivations for data reuse

with a focus on finding trustworthy data. Section 3 describes participants' data reuse following a

generalized model of data reuse. The initial, provisional, and final trust development of

participants is also discussed.

In reporting participants' quotations, I assigned anonymous identification numbers—

from PP01 to PP18 for the public health participants and from PS01 to PS19 for the social work

participants. In the preparation of interview excerpts in this paper, I followed several rules:

- Square brackets [ ] are used for notes that supply information that is not in the

  original words of the participants. For instance, if a participant used the word "it" in

  the sentence to refer to data, I may replace "it" with "[data]" in order to enhance

  readers' understanding.

- Ellipses in single parentheses (…) means I have taken out sentences from the original

  interviews.

- Ellipses without parentheses … indicate a participant's pauses during the interview.

- [Laughter] or [chuckle] indicate the participant laughed during the interview.

To protect confidentiality, all names of the respondents and their institutions and data

repositories were anonymized. Although some datasets were publically available, participants

often mentioned people engaged in the data, (e.g., original investigators or people who were in charge of answering questions about the data who could be traceable), thus these names were anonymized. For the same reason, names of data repositories were also anonymized. However, if participants referred to the name of an institution as a general object or abstract notion rather a specific institution they experienced (e.g., "funding agencies like NSF or NIH"), I included the name in the excerpts.

1. Research participants: demographics and characteristics

In all, 37 interview sessions were conducted with 38 participants (one interview session included two participants). The interviewees were all researchers in various positions (PhD student, postdoc, assistant to full professors, and research scientists), with a mix of genders and ages ranging from the 20s to 70s. Table 6 provides brief information about the participants.

Years of experience as a researcher in the disciplines ranged from a minimum of 2 years to a maximum of 45, with the average being about 15 years. Years of experience with data in research (both secondary and primary data) were also varied, and the participants' experiences with data were not always the same as their years of experience in the disciplines. Some participants had changed disciplines during their careers or came to work with data only later in their research training. Years of data experience in research ranged from 2 to 40, and the average was about 16 years. Except for two participants, all researchers in this study 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 data from government organizations—both federal and state, which are mostly publicly available—and research data from individuals or individual research teams. Eleven

participants had only reused data from institutions; 7 participants only reused data from individual researchers or research teams; and the remaining 20 used both types of data for their research. Data repositories were engaged in the process of acquiring data from both institutions and individual researchers for four participants.

Table 6. *Summary of participants' characteristics*

	Public Health		Social Work		
Rank	Male	Female	Male	Female	Total
PhD student	-	1	1	4	6
Post-doc	1	-	-	-	1
Assistant professor	1	6	4	2	12
Associate/full professor	1	3	3	2	9
Research scientists (PhD)	1	5	1	2	9
Sub-total	**4**	**15**	**9**	**10**	**38**
Total		**19**		**19**	**38**

2. Motivations for data reuse and finding trustworthy data

2.1. Underlying motivations for reusing data

Not all researchers in the fields of public health and social work reuse data for research. The participants in this study had different motivations for reusing data. These motivations were not directly related to their search for trustworthy data but provided an understanding of the reasons for using secondary data for research. In addition, the motivations were often related to

the participants' data-selection criteria.

2.1.1. The potential of secondary data

Several participants said that they would use secondary data only if they thought that the existing data could answer their research questions. Most of them recognized the potential of secondary data—the well-known benefits of data reuse identified in the previous literature. For example, PS05 understood the value of existing data that can be used for "many other research questions beyond the original research… given the vast number of variables" included in the data. "New findings" and "new scientific inquiry" were the terms participants mentioned when they discussed the potential of secondary data. Participants also thought that even original investigators did not fully explore and use the data they produced due to the amount of the data and time constraints.

2.1.2. Cost-effectiveness

Cost-effectiveness is also a benefit of data reuse noted in previous studies. For the participants of this study, using secondary data was also a cost-effective decision because most data are freely available (although some are associated with fees). It helps authors to produce publications quickly if some data collection can be eliminated, which can be a complex process. Time and monetary costs for data collection were recognized by all levels of faculty, as several tenured members noted:

PS10: I really don't do a data collection[;] that's really time consuming.

PS01: I mean, I'm tenured, but numbers do matter, number of publications does still matter, and secondary data analysis usually produce[s] more.

However, it is particularly important for junior faculty who should meet certain expectations to be tenured. Junior faculty in this study were concerned about time, and "always wanted to have, like, secondary data work going on [with other projects] so I can move papers along" (PS09), while some of them tried to submit grant proposals to collect their own data. Secondary data were important for junior faculty career development, as they made it possible for the junior faculty to continue doing research and producing outcomes without funding. PS08 said,

> PS08: I think the easiest answer (…) is it's free, and the funding environment now is so austere and really, really tough, that very seasoned researchers are not getting funded, and so as a junior researcher, it's easier to develop a line of research that utilizes secondary data analysis because it takes…You can still progress in your career without having to always rely on funding.

2.1.3. Large samples

Some participants used secondary data when they needed a large, national sample. This is a distinctive characteristic of quantitative social science data reuse and may not be the motivation for other types of researchers. Having a representative national sample is very important for some social science researchers: when the researcher seeks generalized implications to make a stronger claim, when the researcher uses nations as the unit of analysis, or when the researcher compares the national estimates to the samples the researchers collected themselves. Participants said that it was almost impossible for an individual researcher to collect data that was nationally representative, or there were too many limitations to collect scale data due to the lack of human and monetary resources. Participants knew that receiving grants from funding organizations would make large-scale data collection possible, but they were also aware of the competitive

process for receiving grants as well as the effort required to collect those scale data. For these reasons, participants preferred to rely on secondary data when they needed data from a large-scale national sample.

2.1.4. Education and training

For some participants, data reuse provided great resources for education and training. Using secondary data was a learning opportunity for them, as data reuse exposes "real data, just data that's been cleaned and normalized" (PP15) to the participants without collecting the data themselves. Through experience with secondary data, participants learned "[what] it looks like, how messy it is, how to clean it, how to find the variable, how to work with it" (PP18). Senior researchers and advisors often encouraged the participants to use secondary data to get the necessary skills to use large data sets. Experience with large data sets is important for those participants to make progress in population-based research, which is common in some fields of public health or other research using national samples.

2.2. Motivations for finding trustworthy data

The various motivations discussed above were the drivers of data reuse for the participants. These motivations revealed that the participants viewed data reuse as a worthwhile choice. However, in order to keep their data reuse beneficial and to satisfy their motivations for reusing the data, there was one important condition: data should be trustworthy.

All participants were aware of the significance of data in research and thus the importance of choosing trustworthy data. Participants perceived that data are the "key component of scientific research" (PP08, PS02) and are essential to the "creation of knowledge"

(PS08). Data also "provide evidence to make my claims" (PS04, PS12) and "make the research empirical[ly] based" (PS04). PP11 added that even after one research project is completed using data, the analysis of the project "gives you direction and foundation for the next study."

However, participants were also aware that not all data are trustworthy: "There's a lot of garbage out there" (PP15). She mostly blamed poor study design, inappropriate data collection processes, and unqualified researchers without proper training for the "garbage" and expressed concern about these data leading to biased or distorted results.

> PP15: People have lots of data, and they don't have the right data. And often times what you get out of that is not useful information, and it's very biased because you haven't done the right work at the front end. (…)
>
> A lot of communities or groups that try to write surveys really don't know how to write a survey effectively. (…) Therefore, [they] write a survey that is poorly written, that is biased. Or it could be that they're... Have somebody interviewing people [who] leads them and makes systematic errors or leads them to an answer. [This person may] ask a question but in a way that leads them to a particular response, and those types of things would be not [be] useful at all. Unfortunately, many times people [who] don't have much experience writing surveys don't realize how important writing the questions very carefully [is].

Making a good data selection decision was seen as being central in data reuse quality; PP15 called it "being very critical consumers of data." She thought being a critical consumer of data is a condition for being a researcher in her field—epidemiology—and that everyone should "be

very well trained to be [a critical consumer]." She also thought this skill was something that epidemiologists have been trying to teach and should do more of in terms of "how to" because "as long as [one] follow the rules [to make critical judgment on data], you'd feel pretty good about [the data]."

3. The process of data reuse and trust development

In general, participants' data reuse experiences in the study consisted of several stages: data discovery, data selection (as an ongoing process throughout the data reuse experience, as participants may stop using the data at any point), data acquisition, investigation, understanding and using data for analysis. Participants' data reuse was not a linear process because often they conducted other activities at the same time. For instance, participants partially obtained data files to evaluate data, or participants changed their decision on data selection while interpreting the data when the data did not meet their expectations.

Participants' interview data also revealed that trust played a role throughout the process of data reuse. Data reusers employed a variety of strategies for finding and judging the trustworthiness of data, and they developed trust in the data as they moved through the stages of data reuse. As Figure 2 illustrates, I categorized this process of trust development into three stages. Each stage of trust development can lead to the decision to use specific data (or not).

Process of data reuse

Data selection (On-going process)

Data discovery	Data acquisition	Issue and
Initial selection	Data investigation	problem solving
	Understanding data	Data analysis

Motivation to reuse

Writing manuscript

| Initial trust judgment | Provisional trust judgment | Final trust judgment |

Process of trust development

Decision point   Decision point   Decision point

Figure 2. *Trust development during the data reuse process*

The first stage of trust development is initial trust judgment. This is a process of developing initial trust in the data during the data discovery and initial selection. In this stage, the initial trust was developed before the participant had direct interactions or experience with the data. When the participants developed enough initial trust, they were motivated to move on to the next stage of data reuse, provisional trust judgment, which involves acquiring, investigating, and understanding data for their own research. As the participants had real experiences and direct interactions with data at this point, they developed a provisional trust based on their own investigation, understanding, and use experiences. Participants' initial trust judgment may remain the same or change during this process of provisional trust judgment. Depending on the level of provisional trust, participants can decide whether the data is trustworthy enough to

continue, or they may want to stop using certain data. Even if the participants' trust levels decrease due to a trust violation, as often happens during provisional trust judgment, some participants try to find solutions to these issues. This problem-solving process is part of final trust judgment, in which the participant's provisional trust judgment is confirmed or the trust level is lowered.

This trust development process shows some similarities with those used in past studies, such as the trust development processes discussed by Doney and Cannon (1997), Chopra and Wallace (2002), and Kelton et al. (2008), which together provide a useful framework for understanding data reusers' overall trust development. However, in the context of data reuse, new steps appeared, since the data reusers' trust development aligned with an ongoing process of data evaluation and decision making for reuse. Data reusers developed their trust in order to use (or to decide to use) data for their research, and during trust development, data reusers had multiple decision points for judging trustworthiness and for stopping the use of data that they found insufficiently trustworthy. Data reusers can change their opinions and evaluation of trust after their initial judgment of trust in data, which showed the development of trust as an ongoing process during the multi-stage process of data reuse. It also demonstrated the dynamic nature of trust, which can be increased, decreased, and recovered based on a variety of factors that affect the assessment of the data's trustworthiness.

While Figure 2 provides an overview of data reusers' trust development during their reuse experience, I will explain the details of each stage of data reuse and trust development using the different types of trust that appear during trust development.

3.1. Use of the term *trust* by participants

Although I did not bring up the term *trust* until the end of the interview to encourage the participants to describe their experience in their own words, perceptions and thoughts, many participants used the word *trust* when describing their reuse of data (e.g., PP02 said, "there is trust engaged") or terms that the literature considers similar in meaning to *trust* (e.g., trustworthy, reliable, believable). Most of the participants defined trust as "the ability to rely on" (PS08), "depend on" (PS04) or "count on" (PP15). For example, PP03 said, "I would say trust is your belief in someone or something that makes [it] reliable or it can be counted on."

In general, participants discussed two important characteristics of trust, which were honesty (truth) and good intention. They thought that these characteristics were foundational. Honesty pertains to ethics and morals, and previous literature has explained honesty in terms of integrity (e.g., Mayer et al., 1995; Sheppard & Sherman, 1998). Honesty was mentioned in several participants' descriptions of trust, with a strong emphasis on truth:

- "Being honest about the work they're doing" (PP05), "they are honest with me" (PP13, PS08)
- "To tell me the truth" (PP06, PP12, PS07), "to believe something to be true" (PP09, PP10), "telling the truth, the whole truth, and nothing but the truth" (PP13), "what they are saying is true" (PP05)
- "Isn't misrepresenting any aspect of what I was doing with them" (PP08)
- "Would be either willing to support me by not harming me and by helping me" (PS14)

Good intention was another strong characteristic of trust. Explained as benevolence in the literature (e.g., Doney & Cannon, 1997; Mayer et al., 1995; Sheppard & Sherman, 1998), it is associated with positive intentions or feelings of a trustor towards a trustee. Study participants described this dimension of trust as:

- "Knowing that people have good intentions" (PP05), " […] with good intentions" (PP01, PP02, PP05, PS05, PS06, PS17), "they are not trying to intentionally, mislead people" (PP16)
- "No intentional harm" (PS08), "their willing to not harm me (PS13)
- "Not second-guessing their actions" (PP07), "[not] worrying about what they're going to do now (PS11)
- "Confidence that they made the best decisions they could" (PS13)

Participants' discussion of trust was sophisticated and abstract. How participants decided to rely on something and how they interpreted and determined honesty and good intention in the data reuse context are presented in the following sections.

3.2. Initial trust development during data discovery and initial selection

Data reusers formed their initial trust before they interacted with the data, during the process of data discovery and initial selection. Because it was prior to data exploration or use, a number of social elements surrounding the data affected these initial trust judgments, including the process of searching for and examining the data, the parties that the data reusers interacted with, and any information that the reusers gathered about the data. This section will present the

processes of data discovery and initial selection, and it will discuss how reusers' initial trust developed throughout these processes.

Initial trust judgment → Provisional trust judgment → Final trust judgment

Process of data reuse

Data discovery   Initial selection

Process of trust development

Prediction   Transference
Attribution   Bonding
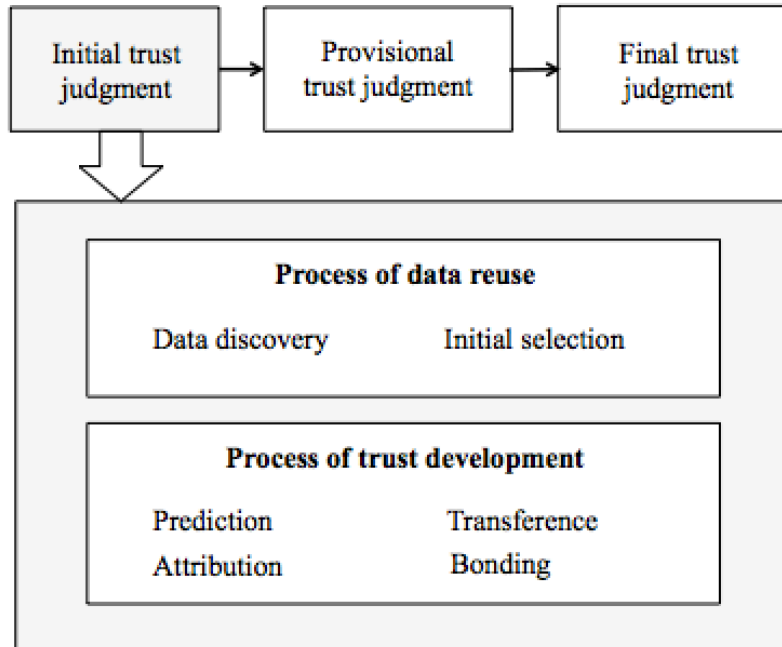
Figure 3. *Initial trust development during data discovery and initial selection*

3.2.1. Process of data reuse: Data discovery and initial selection

3.2.1.1. Data discovery

The data-discovery process was an important first step in data reuse, but discovering data was not always easy for participants; a few noted, "it's not like there is a unified library catalog for entire data sets" (PP17). Thus the participants employed different methods to discover data available for reuse. The major sources and searching strategies of secondary data discussed by the participants were diverse: from literature and education to networking. Trust implicitly appeared during the process of searching for data, though it was not due to the explicit

assessment of the trustworthiness of the data, because the participants had not had any direct interaction with the data at this point. However, the degree to which the data sources were reliable and trustworthy may have influenced the direct assessment of data.

Literature (publications)

The participants often mentioned literature as one way to find available data for their research, whether they were searching for particular data of interest or they just wanted to know what data other researchers had produced and used. Because the literature review was a natural step for most participants when they started to formulate a new research question, the process of becoming familiar with data sources through the literature was also natural, as several participants noted:

> PP09: You kind of get a feel for what's out there when you're starting out learning about the problem. You're going to do your literature search, and you're going to read about the previous studies that have been done to support your hypothesis or get you to the next step (…). So it's always looking to literature first.

> PS02: If you look at the literature of that, you'll see that a lot of the literature, that area uses [a list of the data names], and you will see that they already use [a list of other names of data].

Looking at the literature was also a very useful way for interviewees to learn about small data sets and data collected by individual researchers. It is often difficult to know about small-

scale studies and their data unless one knows the other researcher. Some participants reported that they were able to reuse data from individual researchers identified through a literature review after obtaining their permission. Literature searching was an important source for participants' decision-making on data, and participants perceived literature as a trusted source as the literature had passed a peer-review process prior to publication.

Web searching

Very few participants started data discovery by conducting a web search, and only a few said they "just Googled" something. Those who did search the web conducted a keyword search using Google or Google Scholar. Among those few who had conducted a web search, the majority had a clue about what they were looking for: for example, a part of the name of a data set or data producers. Others used a generic search term, such as "longitudinal health data."

Data websites

Several participants visited the data websites run by data producers, such as government agencies and research organizations (e.g., Center for Disease Control (CDC) or the National Institute of Health (NIH)). These participants had a clear idea about the organizations that produced the data, what data were publicly available, and where they could find the data. These data websites provide a list of available data sets with varying levels of description and documentation. Some of the data websites are well known in the field, according to the participants, and when participants were aware of these sources and they wanted to know what was available, they consulted those websites.

Date repository

Some participants knew of well-known data repositories in social science, such as the Interuniversity Consortium for Political and Social Research (ICPSR). Participants perceived data repositories as "a bank of data" (PP04), "data library" (PS02), or "a warehouse" (PS09). A few participants learned about them in classes; some through an institute run by one of the repositories; and others came to know them through "a natural process, just being in the field" (PP17). The participants said that they started with data repositories for data discovery because they knew "what's in there" (PP17) and "how to do stuff with it" (PP10). Those participants who used data repositories in this study were not novices. PP10 said, "There's a lot of great information in there, a lot of great data sets that you can look at." However, the repositories can be "non user-friendly" (PP10); and novice users were a "little lost what to find, where to find [it]" (PS04). For the experienced users, "it's pretty handy once you know how to do stuff with it" (PP10). While data repositories were useful sources for some, others complained that "[the data repository] is somewhat limited in terms of the data sets they have" (PS14), depending on the research questions or relevant variables, so they did not fully rely on data discovery using repositories.

Formal education

Academic courses are also involved in the data-discovery process. A few participants mentioned having learned about major data sources during their Ph.D. training. Participants learned about the possible sources through informal interactions or casual conversations, as discussed below (e.g. advisors), but also learned from their classes, seminars, and institutes. Not every participant mentioned formal education, but some from each of the two disciplines said

that their Ph.D. program had included a class designated for secondary data analysis or that they

had taken a secondary data analysis class in another department. Those classes offered secondary

data analysis experience and training, and participants remembered that they had also been

introduced to data repositories in order to download the data for the class or other sources they

might use in the future. For instance, PS12 said that a professor introduced the students to a

"good deal of a whole bunch of data sets" as a part of the class, and said, "'these are options for

you.'" In addition, a participant from public health shared that a textbook that he used in the

class contained a chapter describing available data sources, and there were also other books

about secondary data available for students' use.

Advisors and senior researchers

Advisors, mentors, or senior researchers in the team were important sources of data

discovery. They were people who academically influenced the participants by working closely

with them and giving them advice on their research progress. Through a mentor-mentee

relationship, their words carried authority. The participants often noted, "because my advisor

said that," and several actually found and used different sources during their Ph.D. or

postdoctoral training through casual conversations on developing a new research project. For

example, PS15 said, "My advisor, he kept mentioning it and mentioning it, and ultimately I just

Googled."

Sometimes, advisors or senior researchers shared their own data with students or junior

researchers on their team. In addition, they introduced both publicly available and private data

produced by collaborators or peers to their students—the latter would have been difficult for the

participants to access on their own. Participants perceived this to be an important role of advisors

and senior researchers in data discovery: connecting their students to other researchers in their social network. The roles and impact of advisors and mentors are significant in the participants' data reuse experiences as well as trust judgments, as will be discussed later in this chapter.

Social network

Colleagues, collaborators, and other researchers in the field are important sources for data discovery. For instance, PP13 found all of the data that she had used "through word of mouth" through the people around her and was able to meet people who "already knew about [the data]" through collaboration or networking. Several participants underscored this aspect of data discovery: searching through social networking.

Some participants belonged to an established network in their fields, which may not necessarily support data discovery but is important for research (and can be used for data discovery). Being included in the advisor's network was a natural process for several participants; the networks included past advisees, the advisor's colleagues, and colleagues of the advisor's colleagues. PS13 described a network that he joined initially through his dissertation advisor when he was a Ph.D. student:

> PS13: So, a lot of the people who've worked on [the topic] are actually former
> students of his, (…) he holds the network of... There, fairly closely, where we all
> know each other. We all help each other out. We're even having a little
> conference, just like his network essentially, in a few weeks up at [a location]. So
> even if I didn't know them personally, they knew I was to broaden their network,
> and I knew they were to broaden my network.

As people in the network were open to helping each other, PS13 found this network was useful in general for research as well as data discovery. Other participants also commented on the usefulness of having a network, though they were not formally grouped to the level of having their own conference, as was PS13.

Because of the dynamic ties in the participants' social networks, the participants were able to obtain data from both colleagues and from strangers. PP09 talked about both disciplinary and interdisciplinary collaborations as a way of discovering new data as well as a conference as a place to network with other researchers to gather new information about data:

> PP09: I know it sounds like it's a very vague answer, but I think once you're in a discipline and you're working with research teams, you're into finding the stuff that a lot of people are working with, and perhaps through interdisciplinary collaboration you might learn about a new data set. (…) You network with somebody to find a job, and jobs suddenly become available or become apparent to you. I guess I would say it's the same way when you're networking with your colleagues (…) in a scientific meeting.

PP09 compared the process of data discovery to a job search because both involve networking. PS08 also shared a successful experience of actively searching for a particular data set of interest through networking at a conference and a committee meeting:

> PS08: So, we actually didn't have any data related to the research questions that I was most interested in. And so, it was incumbent upon me to find somebody who was collecting data about this population [of interest] and to try and work with

them in order to get my foot in the door, as someone who could become, hopefully, an expert in this area. So, what I did was, I went to conferences and networked that way. I also got on a planning committee for a specialty conference in my area of interest, and through that, I was able to meet a number of researchers interested in my same area and this area, but who had access to data in my area. (…) And then, through that process, [i]f you make the right and the best connections, I was able to actually find quite a few people willing to work on data with me. At least two people came forward and said, "Hey. You know, I think I might have what you're looking for in my data set."

Being in the same network helped them feel safe and willing to share data with original investigators, which sometimes made data discovery and reuse easier for participants. As previous literature pointed out, researchers are still reluctant to share their data for various reasons, but participants found that the network often lowered the barriers to data sharing. PS09 talked about the experience with using "some mildly restricted data" from an original investigator who was not very open about her data sharing through networking. PS09 said, "people are like, 'I own it. I own my data. And you've got to give me a real good reason to allow me to….'" PS09 came to know the original investigator from one of his colleagues in his "network" and believed that the reason for his being able to use the data "that couldn't be networked [reused]," "has to do with his social network." The social network or community of researchers is influential not just in the process of data discovery but also throughout the data reuse experience.

Searching for data through a social network, however, is not always easy for data reusers.

Junior researchers in particular talked about the difficulties of obtaining data without an established network. As PS09 explained:

> PS09: My problem is, being junior, I don't have a great social network. I'm not strongly affiliated with a lab or large research outfit. (…) [The research agencies] have their own data. They collect them and they're sitting there. Half the time they don't know what they're doing, and they're super busy. They want her [the researcher who the agency already knows] to analyze those data [but not me]. I feel like for me, it's more of a process, because people don't know me and a lot of the cases I don't even know what they've got. There may be some information out there. (…) It would be nice if there were more opportunities to find out what data are available.

PS09 was aware of the opportunity that the network could bring him, but without a network, it was difficult to know what data would be available. He also suspected that because he did not have a good network and since the agency that owned the data did not know him well, it would not approach him to analyze the data.

Experience

Several senior researchers in this study who were more experienced with multiple data sets said they were already "pretty familiar with what [was] out there" (PP15). Because they had been working with data for a long time, they often said, "we just all know about" (PP15) or "are just kind of experience[d] in the field and know what exists" (PS13). They went through sources that other participants discussed for data discovery earlier in their career, but they relied on their

experience and expertise now for data discovery, "unlike going to a library catalogue or going to [a data repository], or something like that" (PS13). "Just drawing on years of experience" (PS13), those participants were the sources of data discovery for others; as PS02 said, "I'm the one who [is] giving information about data."

During the process of data discovery, participants did not make a direct assessment of the data, as they were more interested in finding out what was available. Nonetheless, trust was implicit in the sources and methods of data discovery. In particular, when people and communities around the participants became a source of data, those people and communities were usually trusted parties; in turn, the participants conferred a basic level of trust on their data.

3.2.1.2. Initial selection criteria: Relevancy and usability

Participants talked about different initial selection criteria for choosing data for their research. While the initial data selection process was closely related to the initial trust judgment (section 3.2.2.), participants discussed two conditions (or preferences) for data to be reusable for their research: relevancy and usability. Those were not directly relevant to their trust judgments but were essential to the selection process In addition, the initial selection criteria may influence the participants' level of acceptance of trust violations later in their data reuse experiences. Relevancy, the appropriateness of the data for the research project and how well it meets the researcher's need, is a foundational criterion for data reuse, as data reusers choose only data that can solve their research problems. Usability criteria are relevant to the technical elements influencing reusers' ease of using data, which include availability, accessibility, ease of use of the data and its associated software, and the reuser's familiarity with the data type or software necessary to view and process it.

Relevancy

Participants sought data that could answer their research questions or solve their research problems. Thus, whether data met their research needs (e.g., "It had the vast majority of demographic characteristics that I was interested in (PP03)," "It had a best measure of what I'm looking for (PP13)") or was relevant to the topic of research (e.g., "all the variables in the study answered my questions (PP02)") was a common criterion. Only after participants identified the data that were relevant to their research problems did they start examining its other aspects.

Relevance to the research problem means two things to the participants of this study: variables of interest and sample size. The variables can be a measure of interests (e.g., the concept of hope, poverty, and various scales), demographic characteristics (e.g., race or ethnicity), or a population of interest (e.g., LGBT (Lesbian, Gay, Bisexual, and Transgender), over 65). Depending on the research problem, participants checked whether the data was nationally representative or whether the sample size was large enough to run a statistical analysis. Although an adequate sample size was important for some types of analysis, not all data met this criterion. PP14 talked about "encounter[ing] quite a few times with data sets that I've been very, very interested," but being unable to use the data since "they have such a small number of group[s]."

Usability

Participants expressed a preference for data that were easily accessible and freely available. As discussed in an earlier section, one motivation for participants to use secondary data was saving time and money. For example, for PP02, the rule was to use publicly available data that was free and easy to use, and for PS15, the time and effort involved in obtaining data

were important criteria:

> PP02: Proprietary is also important. I, as a rule, don't spend a lot of money on
> purchasing secondary data sets.

> PS15: It's one of those cost-benefit analyses kinds of things, right? (…) There's a
> good number of hoops to jump through to get access to [the data] (…). And it can
> be one of those things where if it's too difficult of a process or too lengthy of a
> process (…) Cause sometimes, people have very short timetables they're working
> on for whatever reason, it might not be worth it to invest all that time and energy
> to gaining access and getting set up with the data.

Even data that are publicly available may include portions that are restricted, requiring a special application. This extra step may not entirely prevent use of the data, as "it becomes that trade-off," and the decision depends on "how important it is to you, to look at Y, when you're really interested in X, and how much time and effort you are really willing to put into looking into [it]" (PS15). However, some participants expressed a strong preference for finding data that are fully publicly available to save time and money.

Data formats and software were other elements that influenced the usability of data. Several participants reported difficulties with formats, and others shared difficulties using unfamiliar data software or analytics programs. For instance, PP01 struggled to download some data because of the format:

> PP01: There're also data sources that are incredibly labor intensive to get the data

downloaded and functioning. [These data] have different data formats (…). [I]t

just becomes a huge headache.

PP11 was also unable to open data because "there's a kind of, some process involved." She was not sure of the format of the data or if she needed a special program to open and run the data. She needed "some assistance from a statistician or other experts," but she ended up abandoning the data, thinking, "It's just too much," without any institutional support or data services available to her. PP12 had tried to use some data from Europe: "You can't just use regular SPSS [Statistical Package for the Social Sciences] for that data. You have to use programs that are set up for [it]" (PP12). She had to use the analytic program designed for the data format but was not sure if the new software program was worth her time and money. Thus, participants sought data that were "at least transfer[able] pretty easily" (PP05), and they were "probably a little reluctant to use [some data] because you do have to use [special] statistical software…" (PP12), depending on the research support level at their school or institution.

3.2.2. Process of trust development: Initial trust judgment

Data reusers made their initial trust judgment before having a direct interaction with data. At this stage they may not have developed full trust in the data, but they perceived that the data were trustworthy enough to continue to the next step where they would have a direct interaction with data by acquiring, investigating, and understanding them. Without having this actual experience with the data, the participants developed their initial trust though the processes of *prediction*, *attribution*, *transference,* and *bonding* (Figure 3). These mechanisms shared conceptual similarities with the steps of trust development discussed in the previous literature.

3.2.2.1. Prediction

When data reusers had previously worked with data, they already had formed trust in those data, and were therefore willing to use them again. Both Doney and Cannon (1997) and Kelton et al. (2008) named this type of trust formation as *prediction* and explained that *prediction* is based on trustees' past behaviors or the data reusers' past experiences with an information source. The participants who used the same data again 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), and because "it made [a] really easy transition because my previous project kinda familiarized myself with those data" (PS08). This reflects the efforts data reusers should spend on understanding new data for reuse and on judging the trustworthiness of data. PS12 remarked, "Continuing to analyze off the same dataset [that I trust] is a brilliant idea because you minimize that learning curve" although he had used several different data sets.

When data reusers found data sources (original investigators) to be trustworthy from past experiences, they tended to trust new data from the same sources. PP15 said, "it's a matter of working with these people over time (…) and consistency in the experiences." PP01 also expressed her trust in data from a certain institution that she had previously used:

PP01: I wouldn't say that I automatically trust the data produced by the [institution's name]. (…) But I've been doing this [research] for 15 to 20 years at this point and working with [data from the institution], and in the course of my work life, I've never had an issue with their data, and the people [in the institution] have worked there professionally. (…) I know that [the institution] is a

trustworthy source, I know I can trust their data.

These cumulative past experiences built strong relational trust between data reusers and the sources of data (original investigators) and the data itself.

Sometimes, participants expressed a strong preference to work with data that they had previously worked with or data from original investigators whom they already trusted:

> PP15: I just don't have the time to focus and learn and assess a [new] data set. (…) I don't look for a data set that answers my research questions. I tend to know data sets that I'm comfortable with and focus my research questions on what they tend to cover.

For PP15, working with trusted sources was the first priority when selecting data, unlike the other participants, who set up the research problem first and then looked for data. She focused on three or four data sets that she had trusted from past experiences and chose research questions they could answer.

3.2.2.2. Attribution

Trust can be based on data reusers' rational choice and judgment, stemming from credible information available concerning the data. Previous trust literature has named this process of developing trust based on observable evidence *attribution* (Chopra & Wallace, 2002). In this study, participants discussed three types of evidence that influenced trust development: existing evaluations of data, the competence of the original investigators, and the intentions and the ethics of the original study that produced the data.

Existing evaluations

With a lack of direct interaction or experience with specific data, participants first sought evaluations of the data and based their initial trust judgment of the data on that information. This was a rational process to develop trust by seeking confirmation from various sources that the participants already trusted. Peer-reviewed publications were very useful sources on which several participants relied. PS10 considered publications as "one step further to establish trust" in data, and PS13 said it is helpful to know "if there's something that's already published with the data in peer-reviewed journals." Because participants understood that a study "goes a long way" and "[has gone] through a rigorous review process" (PS09) in order to be published, they respected the role and expertise of reviewers. Participants already trusted reviewers and the review process, and data that had undergone peer review had passed the "bar" for acceptance in research communities, with "no problem" (PS12).

> PS07: [T]here's a level of peer review that you can trust in terms of thoroughness, and an expertise among those who are reviewing it and to weed out junky stuff.

Thus, the results of publishing (using the data) demonstrated that "the data were sufficiently trustworthy to be worth publishing in an important journal" (PP13). The fact that someone published an article using the data also bolstered the participants' confidence that a data product was "acceptable through the normal channels of scholarly activities" (PP17). PS07 said, "You know that you can have confidence [about the data], and if you are also able to do some sort of decent analysis that you'll be able to kind of go into that level, a very reputable high tier conference or publication."

The number of times data had been used was another data quality indicator. When data were "widely used and widely cited," the data "must be good" (PS12) and "widely trusted by others" (PP17). PP17 commented, "That's evidence and a component of trust." Due to the lack of data citation standards implemented in the fields at the time of the interviews, participants checked the numbers of publications that used data and the number of citations for those publications. Often they informally checked how many others had used data, including colleagues and other researchers in their field.

> PS04: Most of the data that I looked at was already known to a lot of the
> colleagues I work with. And that already told me that this is a good enough data
> [set] to use for this research field.

For them, peers' use guaranteed the data quality as an indication of the data "[being] well respected and validated" (PP05).

The fact that original studies were funded by either government or non-profit organizations was another indication of peer-evaluation and communities' acknowledgement. Participants perceived data funded by organizations like NIH and NSF as "more reliable" (PP11) and "trustworthy in a sense" (PS02) because "if [the study] received grant funding, it [was] vetted by a group of their peers" (PS08). Participants understood the process of competing for grants, and thus, being funded was a recognition of outstanding research with appropriate methodology that "makes [one] feel better about [the data]" (PS12). PS07 also talked about federally funded data: the "government had spent five million dollars on the study to make it. (…) and g[iven] the researchers' appropriate and sufficient resources to do a study well," nothing "[the government] wouldn't just waste that much money for nothing," which made her evaluate

the data as "good data." Although participants acknowledged that not all funded projects created high-quality data, in general, participants believed "if data comes from a grant-funded project, you're more likely than not to get a high-quality data set" (PS08).

Competence of the original investigators

As identified by previous literature, participants in this study considered several aspects of the original investigators in developing their initial trust in data starting with the competence of the original investigators. 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. While there are other ways to check the competence of the original investigators, such as their reputation (see section 3.2.2.2.), some participants suggested membership in a Community of Practice (CoP) as a way to check their competence.

Shared concerns, experiences, and practices are the characteristics of a CoP, as defined by Wenger et al. (2002). In particular, participants wanted to make sure 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." PS12 echoed, "I'm gonna check the people, like whether they are trained like me, like do they have Ph.D.s?" However, training meant more than simply holding a Ph.D. PS12 continued, "I mean, but that's not the sole criteria." Training can be common in a discipline or sub-discipline; a few health researchers noted that they shared a certain approach to data and data education as "epidemiologists" (e.g., emphasis on using national sample data). Training can also be interdisciplinary, such as training in quantitative 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 original investigators' home departments as well as the departments in which they had been trained (which implied how they had been trained), checking the main methodology the original investigators had used in past research, and identifying their research interests (which can imply a core methodology), participants were able to ascertain whether the original investigators were part of their CoP. As previous literature on CoPs has argued (e.g., Hislop, 2004), there were already trust-based relationships enhanced by a consensual knowledge base and shared identity that helped members to share knowledge and accept opinions of others within the CoP.

Intentions and ethics of original studies

The data producers' underlying motivations and intentions for conducting research and producing data were also important for the participants' initial trust judgment. Many participants discussed the importance of the original investigators' research ethics and integrity, meaning that researchers should collect and manage data ethically, and refrain from manipulating data. However, the ethics of the original study and its investigators were not always easily verified or evident in the study itself. In such cases, the participants often used proxies in assessing research ethics.

Participants tended to develop their initial trust in data if those data were from government organizations or nonprofit organizations, or if the data collection was funded by nonprofit or government agencies. Participants compared these institutions' intentions toward research to those of private, for-profit corporations.

PP04: [S]o, questions like, who funded them? If it's a pharmaceutical company, I

would be very skeptical, even if they are very well known about using their data.

And if it's one of the NIH Institutes, I would be very comfortable using their data,

and if it's funded by someone who doesn't have any vested interest.

PP04 was concerned about conflicts of interest and data being created for a specific purpose or becoming skewed if the study had private or for-profit funders. PP04 emphasized the purity of data that was created for the purpose of scientific knowledge and the public good, especially when the research was funded by non-profit research organizations.

Checking the ethics and intentions of the original studies could be more difficult when the participants dealt with data produced by individual researchers who were unknown to them and had no research funding. (The mechanism to check ethics and develop trust in data from known individual researchers with in-person relationships was different. See section 3.2.2.3.) When participants did not personally know individual researchers, checking their social identity was an alternative. Participants wanted to know who the individual researchers were and ensure that they were actually researchers. Having same social identity as a researcher generated a sense of kinship and allowed for foundational trust. There was a high propensity to trust members of a group with a shared identity. This is known as "in-group favoritism" (Aronson, Wilson, & Akert, 2010, 2010).

For instance, PP16 talked about trust in individual researchers and their data, integrity, and good intentions "not to mislead anyone" after checking the social identity of individual researchers. PS02 attested to the academic integrity of the researchers after checking their credentials as researchers and as professors:

PS02: They are assistant professors for tenure track; they are working at the university. So I think that because they are already professors, and they are already scholars. (…) I assume academic integrity, that I assume that they are not even giving me that data [laughter] that they've made, they cooked up.

In this case, the social identity of original investigators—specifically their institutional affiliation—guaranteed their membership into research and academic communities at large and thus gave them the same level of ethical leverage and academic integrity.

Participants' in-group favoritism and their trust in the original investigators' ethics and integrity were supported by the broader level of social trust, which was invested more generally in research communities and researchers. Participants discussed their own commitment to research and their pride in being part of a research community. PS17 said, "We are doing science, scientific research, (…) for public good." PP18 believed research "[is] conducted in an ethical and credible manner." PS18 mentioned a "code of ethics" to which all researchers, including himself, should comply:

PS18: I think we, the researchers, all know things like honesty, objectivity, integrity… and have to conform with them in research, you know, being accountable to the public, and so on… I think these are social responsibilities of researchers.

As participants understood ethics, integrity, and social responsibilities of researchers and "[being] confident doing good research" (PS18), they had the same expectations for other researchers to conform with socially accepted behaviors as researchers.

PP09: [Y]ou don't make up data for 30,000 people a year. Neither anybody.

Nobody's putting up arbitrary stuff. Nobody's got that kinda free time on their

hands.

Based on the trust in other researchers, the entire research community, and the general practices

of research, participants believed "the people who are collecting the data don't have some

ulterior motives to screw up the data" and thought "it's in their best interest to do a good job

collecting the data and to have a good understanding of the data" (PP09), even though some

people "can be jerks and can be cheap" (PS12).

3.2.2.3. Transference

Doney and Cannon (1997) explained that trust is developed and transferred from other

parties, and that data reusers also developed their initial trust based on other people's perceptions

of the data's trustworthiness. This is the process of *transference*. Participants based their trust

upon reputation and recommendations. They accepted both as credible information and a

research community's positive acknowledgement and had high expectations of data with a good

reputation or that came from reputable original investigators. This view was also based on the

participants' rational choice and thinking, which brings calculus-based trust.

Some participants relied on a person's reputation to judge the trustworthiness of data or

the sources of data. PP17 described data he used as "the" data, and also acknowledged the parent

study's reputation, saying, "The study is considered 'the' study." Because the study was known

for its "terrific study design" with "good sampling and measurement," PP18 "felt good about the

data." For both PS04 and PS02, data from "renowned scholars" or "reputable" organizations

meant trustworthiness:

> PS04: I haven't met any of the researchers directly, but they were already
> renowned scholars in the field. So, most of the research with most of the articles
> written by them have high citation numbers. (…) [T]his was my decision whether
> these people were trustworthy.

> PS02: [The data] is collected by a reputable and ethical agency like NIH.
> Basically that I don't question [that] because they are all really trustworthy
> organizations. (…) I blindly trust the data that it's from a reliable [organization]. I
> don't question the data.

Further, reputation implied positive characteristics: Reputable data implied a rigorous original study; reputable original investigators were competent and had integrity, while a reputable organization had the internal capability to conduct research. 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 like NIH or CDC made "[me] lay the responsibility with that organization to collect the data properly (…) and as best quality."

For data produced by institutions, such as government agencies (e.g., CDC, NIH, or USDA), the participants recognized the reputation of the institution as one whole party rather than individuals within the institution, which provided sufficient credentials:

> PP02: It's [the agency's name]. It's not like it's Bob down the street.

Participants relied on the institutions' reputation, which "wasn't necessarily bas[ed] on the

reputation of those [individuals] that I had seen as authors of the [data]" (PP03). Participants

pointed out the reasons for not checking the individuals involved in creating the data at the

agencies or organizations. Because the agency's reputation had been built through other data

reusers, it was natural for participants to accept the reputation of organizations known for "their

quality data" (PP07). At the same time, participants believed in the organization's competence

and that they "physically employed the trained people that are data collectors" (PP01), which

made it unnecessary to see who worked for the project to create the data. The nature of trust

resembled the concept of institutional trust identified in previous literature: trust in the

institutional structures, properties and competence (e.g. McKnight et al., 1998; Rousseau et al.,

1998; Sitkin & Roth, 1993). When participants did not know the people responsible for data

collection and did not have a direct relationship with members of the institutions, this

institutional trust could substitute for interpersonal trust (trust directly in the individuals who

produced the data).

For data produced by individual researchers, the reputations of individuals—particularly

the original investigators—was more important for participants than the reputations of

institutions. PP18 remembered the original investigator of the study as "a well-recognized expert

in the field," and everybody knew that the original investigator and his research "were really sort

of groundbreaking." This gave PP18 "a lot of respect for him and his research," in addition to a

"certain amount of trust" in the data.

However, reputations of academic institutions to which the individual researchers

belonged (institutional trust) could still enhance the participants' trust in data. They also

substituted for interpersonal trust when participants found data from an unfamiliar source

without a notable reputation, such as a junior researcher. PS02 referred to the institutions where

the individual researcher worked at a "very prestigious organizations (…) with established people," and PS12 was "probably influenced a little bit by the reputation of that institution" because it was "a very high-powered, good research university," which made PS12 think "[the data] must be good." Still, institutional trust in this case only supplemented the participants' trust development on data, and the reputation of the researcher's institution "only gives a positive bias" and "doesn't lead [to] a negative bias," according to PS12: for instance, data from a non-R1 university did "not necessarily mean 'Oh well, you shouldn't trust that'" (PS12).

Direct recommendations on data from colleagues, advisors, or collaborators, were also part of the overall reputation that participants considered, as it was the opportunity to learn about other people's perceptions of data trustworthiness. This is important because participants "respect[ed] what other people have felt about data" (PP17). Sometimes, junior researchers sought a recommendation from senior researchers or their advisors:

> PS04: I asked people or colleagues or advisors about data that they knew of that is
> up to date and that is relevant to my research questions. And this was one of the
> data [sets] that they mentioned: Ph.D. colleagues who [have] already graduated
> and are working as faculty. And, some of them have used this data or some of
> them just heard that this is good data to use for research. And my advisor also
> suggested that this is good data and you can use it well.

Often, recommendations from a party with an established authoritative relationship influenced the participants. PP13 planned to use data for a project when she wrote a proposal because one of her colleagues had used it for another project and recommended it. She explained that "we know him really well, and we trust his word about what the data can do." PS07 said, when a senior

person "who's a well-funded, well published, associate or full professor who [is] familiar with data analysis and research" advises him, he "typically take[s] it pretty seriously."

For some participants, checking the reputation of data among the researchers around them was equally important to verify the reputation in the field, as they had received advice against using certain data. PS07 talked about "having people tell me, 'Hey, don't get involved in. (…) The data [are] a mess,'" and PS08 also had a "close colleague [who] warn[ed] me from something [about data]." Usually a negative recommendation was made privately among researchers who were in a close relationship, as PS12 described:

> PS12: It's not like shouting in the middle of the conference. (…) I mean, we had colleagues, and I can't go into a lot of details, but I have had colleagues who told me about the data I should be worried about, and I wouldn't use or recommend working with them.

3.2.2.4. Bonding

Data reusers also developed affect-based trust from their emotional connection with data and the parties behind the data, mostly original investigators. Kelton et al. (2008) and Chopra and Wallace (2002) proposed calling this type of trust formation *bonding* and explained it as the process of emotional development in a trustor-trustee relationship. Since it took place before data reusers had worked with the data, this affective trust was developed from the interpersonal relationships with people relevant to data (interpersonal trust) as well as cumulative experiences with them (relational trust). This established relationship and social tie built a strong trust between people, which made the data reusers rely on parties whom they believed to be

108

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'."

Participants tended to accept data that came directly from "the interpersonal connection" (PS13), "not some anonymous person" (PP10), because "familiarity breeds confidence" (PS12). This was important for PP04 when searching for data. A close relationship with a data creator allowed him to trust the data because he already knew enough about the original investigators.

> PP04: I will ask only those people whom I trust that I have a working relationship with them because now I have been in this field for long enough to know whom I can trust and who is the person who wouldn't do anything unethical.

The same trust inheritance occurred when participants took recommendations from their colleagues or made their decisions on data from their colleagues' use of that data.

> PS04: Most of the data that I looked at was already known to and used by a lot of the colleagues I work closely with. So, I didn't have any issues with trusting.

> PS09: I think with a lot of these data it really is... I mean I know this is gonna sound strange, like when you meet someone, you trust them based on their friends at least initially. You know what I mean, like, a friend of yours introduces you, "oh this is my friend," you automatically [think], "if this person likes this person, maybe I'll like them too." So if someone says to me, "I've worked with this data, you can trust it," then I'm inclined to trust it.

Because participants were already in trusting relationships with advisors, mentors, or senior researchers in a team, they tended to trust their mentors' opinions on specific data almost without reservation. When participants described searching for and selecting data from their advisors or mentors, they usually did not say more than "my advisor suggested" (PS04) or "just through conversations with my advisors" (PP17). This mentor-mentee relationship conferred authority upon the advisors' work as "we know [the advisor] really well and respect his work a lot," which led PP12 to "trust his word about what the data can do" (PP12).

Sometimes, this affect-based trust judgment was built upon the participants' belief in the trustees' competence and their integrity.

PS04: They are trustworthy scholars. I've known them already, long enough to know whom I can trust, who is the person who wouldn't do anything unethical. They are sound researchers. So that's one of the bases, one of the criteria.

PP04: I usually only start considering data that I already trust and that I know it's been collected by a good source. (…) I felt really comfortable that I could trust the data because I knew the person really well, who had collected it.

As PS04 and PP04 illustrated, both of them already had a relationship with the original investigators and assessed them to be "sound researchers" and "a good source." This association made their data trustworthy. This calculus-based trust was grounded in the participants' emotional attachment to the researchers, which increased the participants' belief in the benefits of their data.

110

As already noted, several social elements around data influenced the development of reusers' initial trust in the data, through a process of prediction, attribution, transference, and bonding. While a low level of initial trust may not prevent further investigation of data, as some reusers might still want to investigate the data by themselves despite their suspicions, initial trust played an important role in the early stage of data reuse by leading reusers to the next stage: acquiring and investigating data with a high level of trust in the data. In conclusion, initial trust is often formed before data reusers actually use the data. This initial trust cannot guarantee that reusers will ultimately trust the data until they fully explore and use it through the provisional and final trust judgments.

3.3. Provisional trust development from data acquisition, investigation, and understanding data

The process of establishing a provisional trust judgment aligned with the data reusers' own acquisition, investigation and understanding of the data. The process of provisional trust judgment development through in-depth investigation and coming to an understanding of the data was "the time-consuming part" (PP13, PP15), although a few participants said that investigating and understanding the data "was not hard" (PP11) for those who had years of experience with similar types of data or those who went back to the same data multiple times. For many others, especially those who worked with new data, it was a "difficult learning process in [the] very beginning" (PP14), which required "a lot of trial and error" (PS15). Because every data set has unique features, PP15 stated, "You really need to understand [it] very thoroughly," and compared it to learning a new language:

PP15: People have come to me and just showed me a data set and said, "Can you

get me this, this and this out of it?" And you can't do that, you've got to really

know the data. You've got to understand it, and it takes a lot of time to kind of

immerse yourself. It's like somebody speaking a language and just saying, "Can

you just learn Italian in a week?" You can't do that; it takes time, and it takes a

real understanding of what goes into the data set and what's behind it, at least for

me. So it's a fairly complex process to learn a data set; well, that's why I say that

we've really become experts in a few data sets.

PS15 noted the time commitment in understanding data and said that data reusers probably spend

"much time trying to figure out the structure and the questions." Thus, participants needed to

"immerse" (PP15) themselves into the data.
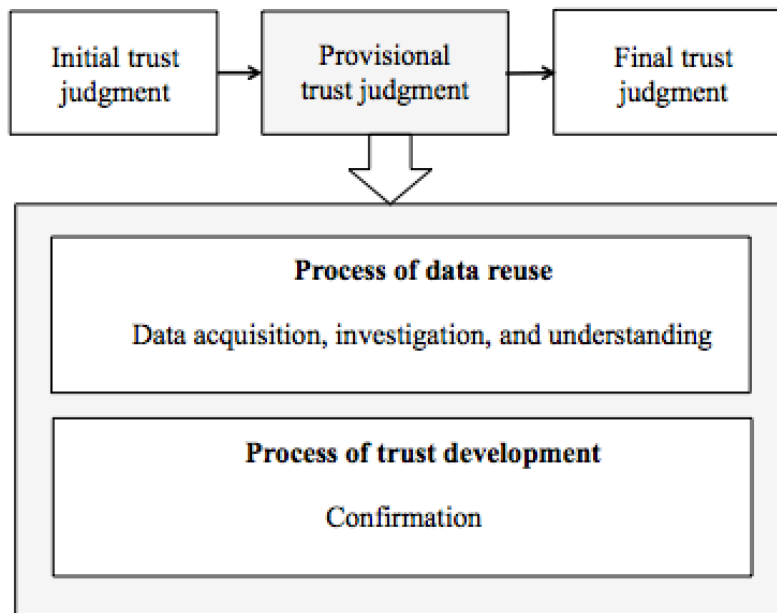
Figure 4. *Provisional trust development during data acquisition, investigation and understanding*

3.3.1. Process of data reuse: Data acquisition, investigation, and understanding data

3.3.1.1. Data acquisition

Participants in this study acquired data directly from the original investigators or from data repositories. Different methods of acquiring data were discussed by the participants, depending on the data's source (see Table 7).

Unlike the data acquisition from individual researchers that involved personal interaction with the original investigators, less human interaction was involved in acquiring publicly available data (either directly from institutions or from data repositories), when the data were freely available without any restriction. Participants usually downloaded the data, with or without registration, and most participants evaluated the process as "pretty easy" (PP06), "pretty straightforward" (PS10), and "just have to click and download" (PP13).

Table 7. *Data acquisition methods*

| | Original investigators | | Data repositories |
	Institutions	Individual researchers	
Download from websites	X	X	X
CD/DVD	X	X	X
External drive	X		
Server sharing		X	
Email		X	

When participants wanted to access "confidential data" (PS01) or a portion of restricted data, they had to "go through certain processes" to "satisfy all the requirements" (PS01). Sometimes, participants had to persuade the institutions or data repositories of "the importance of having access to the confidential data (…) and why it was necessary for the project" (PS03). When participants liked to work with data from individual researchers, the relationship between

113

them may have eased the process of obtaining permission, and PS09 reported working with "some mildly restricted data" due to the personal relationship with an investigator who was originally reluctant to share her data.

Not all participants had to sign a use agreement form, particularly if the data were publicly available and de-identified. When participants received data from individual researchers with whom they were in a close relationship, there was no use agreement form as the acquisition process "wasn't really that official" (PP08), except for those cases when some participants reported that they had signed a form. The use agreement form usually stated that the participants "should not try to profit from the data, […] should use it for the purpose intended" (PP05), and "should not do anything unethical with the data" (PS08).

Participants also described what was included in the data package that they downloaded or received. Although there were some variations in terms of how participants referred to the content, participants listed similar elements in the data packages. The files in the data package that participants mentioned during the interviews were as follows:

- Data files (either entire data sets or several variables of interest)
- Study instrument or questionnaire
- Code book or data dictionary that defined and explained variables
- Other documentation or data notes that explained sampling, weight, data collection, or background of the study
- A list of publications that used the data

There were differences in each participant's recollection of how those files were organized, how

they were prepared, and how clean they were. In addition, sometimes data from individual

researchers did not come with well-prepared data packages or documentation associated with the

data files. These differences in documentation affected the participants' understanding and use of

the data, which also influenced to their trust judgments (See 3.3.2).

3.3.1.2. Data investigation and understanding

Once data reusers acquired the data from either the original investigators or data

repositories, they began investigating the data and tried to understand them. Investigating and

understanding data are crucial in data reuse in order to transfer the contextual information about

the data and the original investigators' knowledge to the data reuser. The investigation and

understanding of data differed slightly among participants depending on their workflows or

practices, but in general, participants discussed several common procedures they undertook.

They tried to understand contextual information about the data by either going through all

relevant documentation or by contacting the original investigators (where there was no or

insufficient documentation) and also by reading articles published using the data. During this

process, participants simultaneously investigated and checked different aspects of the data, such

as validity and reliability. While some participants focused on examining and understanding the

data, others extracted their variables of interests or created working datasets by merging different

variables or by merging data from different sources during this process. During the reuse phase,

they sometimes ran into unexpected problems that sent them back to the documentation.

Reading documentation

Reading documentation was one channel of understanding data, and several participants

noted the importance of documentation. Documentation is "the key" in data reuse, and PS14

said, "[It] is crucial. If there's not adequate documentation then we don't know what's going on in the data." Participants read the documentation and looked for contextual information that enhanced their understanding of the data. The level of thoroughness may depend on the participants' experience, subject expertise, or tacit knowledge, which cannot be captured in this study, but all participants wanted to know how a study was designed (including sampling design); how the data were organized and structured; how the data were collected (often with more detail, such as how interviewers were trained); what changes had been made to the data and cleaning processes used on the data; what the variables meant and how the variables had been recoded; what and how measures or scales were used (if applicable); and how the original investigators had analyzed the data. The quality of documentation was significant for further use of the data, and poor documentation could deter participants from using the data:

> PP09: [If] it doesn't contain enough of the information you're looking for, you don't use it.

> PS01: If documentation is poor and it doesn't seem to have the information I'm looking for, that's probably not worth to move on.

The comprehensiveness of the documentation varied from the participants' perspectives. In general, participants evaluated documentation of institutional data as "extensive" (PP01), "outstanding" (PP09), "pretty thorough" (PP10), "very comprehensive" (PP13), "pretty complete" (PS03), "phenomenal" (PS15), and "easy to work with" (PP15). They were satisfied with the documentation, except for some data with "sloppy documentation" that "doesn't make sense" (PS09). Documentation from individual researchers can be "easy to follow," "very

thorough" (PS08), "very detailed" (PP10), "easy to follow and straightforward" (PS10), and "excellent" (PS08, PS19), but data from individual researchers also had documentation that was "poorly described," "not very well put together" (PS13), and "difficult to understand" (PP03) with a case of "basically no documentation or codebook" (PS15).

While it is hard to say all participants considered one better than the other, several participants had strong views regarding two types of documentation: documentation associated with publically available data and documentation from individual researchers upon request. Some participants (e.g., PS15) preferred documentation from publically available data because it was ready for reuse and was more complete.

> PS15: So I think the [data from a government institution], the datasets that are really well-known and used often, you're not gonna probably have that problem where there's a lack of information because that's used a lot by many people. I don't think it's the case in the smaller datasets that people (…) haven't used much, [when individual researchers] would like something done with it and they offer it to others.

However, a few participants (e.g., PS07) expressed a strong preference for documentation directly from individual researchers, claiming that sometimes the documentation of publicly available data barely met the requirements of funding agencies and was not adequate for the purposes of reuse.

> PS07: [Documentations of publicly available data], it's user-friendly, right? But sometimes they're uploading just... It's usually they're probably assigning it one of

117

their research assistants and saying, "Hey. We need to get this thing done fast."
They just upload whatever documents they have there, so they may not
necessarily be cohesive. But if you're using the data [and documentation] that the
investigator, he or she, has used to publish, it's usually way richer, [of a] higher
quality, understandable, etc.

Both views were valid since they were from the participants' own experiences. However, the
issue was not the type of documentation the participants worked with but the quality and
preparedness of the documentation. As long as the documentation was found to be detailed,
easily understandable, and well-organized, it was useful whether it was from an institution, from
publicly available data, or from individual researchers.

Getting information from original investigators (individual researchers)

Only a few participants said, "there was no documentation" (PS15), and a few had
worked with a minimal amount of documentation—at least a description of the original study or
"some sort of annotation about data" (PS17). The cases with little or no documentation were
generally the data from individual researchers or individual research teams rather than from
institutions or data repositories. While participants did not always know why there was no
documentation, a few assumed that the individual researchers had not expected other researchers
to be interested in using their data.

For cases with little or no documentation, some participants already had a previous
relationship with individual researchers, so they were able to understand the data through direct
interaction with the original investigators. Further, a close relationship with individual

118

researchers enabled participants to use data that were less well-documented, as they had access to the individual researchers. However, relying on the individual researchers for information about the data and understanding was not always easy. PP03 tried to understand data from scratch:

> PP03: When I actually start getting into the data sets, [I found that] it takes just an extraordinary amount of emails or conference calls with the scientists involved [in data creation] to just understand how the data was collected and potential problems that may have popped up when the data was collected. Protocol deviations and things like that.

In contrast, PS18's experience was "very smooth" and it involved "just a few meetings with [the original investigator], I think it's actually one or two, then I can always call her for more information." The differences in these experiences may be due to the participants' experiences, tacit knowledge, work styles, or the closeness of the relationship with the individual researchers.

Reading publications

Participants also went through publications, articles published by the original investigators, or articles published using the data. Publications were already discussed as an important factor in participants' initial trust development. Participants used the publications to aid their understanding of the data and their use for their own research: "there was a lot of information beyond the actual codebook" (PS07). Typically, these publications had a paragraph with brief information about the data and its strengths and significance. This gave "a nice summary of the data" (PS18) to the participants, saving them hours of research time to

understand the background of the studies.

In addition, original papers instructed the participants regarding how the original investigators had used the data, which several participants found useful. Participants also looked for other pieces of information "that would complement my findings" (PP06). They took a special interest in reading "the limitations of the data from the people who already experienced using it" and wanted to make sure that they "didn't really see any huge limitations for the particular purpose for this data set" (PP06). As discussed earlier, because participants respected publications that had undergone peer review and respected the authority of those publications, they believed that the quality of the data discussed in the publication "should be [at an] acceptable level" (PP06)—not just according to the participants but also as evaluated by other researchers in the field.

Checking data

While understanding the data by reading documentation and publications, and interacting with the original investigators, participants continuously checked different aspects of the data. They engaged in a close examination of the data to see where they would spend "the bulk of [their] time" (PP05). Sometimes participants tested the validity, reliability, and missing data. Some participants examined the data for consistency, transparency, and quality of documentation. What participants wanted to know was directly relevant to their trust judgments, which will be described further in section 3.3.2.

3.3.1.3. Common difficulties

Searching for information

One common issue for participants was the difficulty they had in finding information they needed in either data packages or associated documentation. Several participants had used data that were "much more difficult to understand" (PS08) because they had a hard time finding the information they needed. PP01 said, "the data make me work really hard to get the information I need, and there'[re] just times where I wasn't sure it's worth it." In this case, the difficulty was not from a lack of documentation or information but because of the organization of the information. Another participant noted that one data set had "separate coding sheets that are not in their data dictionaries" (PP12). Neither PS15 nor PS14 were able to find items that were supposed to be in the sections they consulted, and they initially thought "maybe it was excluded for some reason, but it turns out it was elsewhere." All three had to "seek out the information" and figure out where the information could be found in the entire data package "to add it back in" (PS14), which they found time consuming:

> PS15: They have codebooks available, but it's very time consuming to really sit there and wade through it. And it's one of those things that you have to do due diligence to make sure that you don't overlook information that you want, and say it's not there when it actually is there.

Understanding codes

Sometimes the participants just did not understand the labels that the original investigators used: "I don't exactly always know how [they] did label [the data or the variables]"

(PP12). PS08 said, "The codes were not intuitive" because "you can't make guesses about what you think things mean." Although there were codebooks, codes and labels that were readily understandable could have saved time and lightened the workload. PS08 recalled working with inappropriate codes, "spending much unnecessary time to understand the data," but she still did her best because an incorrect understanding would only give "3,000 observations of numbers that are meaningless."

Insufficient documentation

One of the greatest challenges participants discussed was insufficient information. A few participants had worked with data from individual researchers, accompanied by minimum information in data packages, occasionally even without codebooks. PS15 just received a "paragraph about the methods, and had a paragraph, [that] maybe talks about the population" but did not receive "a codebook and a lot of demographic questions that they've been asked." PS13 also received "only the major data files, there really isn't a codebook," and any other documentation was "just very poorly described" with "some sort of abbreviation that you're not exactly sure what it is." Insufficient information made a participant feel that he "kind of ha[d] to know by myself" (PS13) and "assume it was coded this way, but I could be wrong" (PS15). Both of them eventually used these data sets because they trusted the credentials of the original investigators and thought that the poor documentation was not relevant to the original investigators' intention, but "they just didn't document more than that" (PS15), not even considering the possibilities of data sharing, which PS13 thought was "very unfortunate." Because of the lack of information and poor documentation, both participants spent time figuring it out by themselves, but they would have liked more information about the data.

PS15: It was enough for publication purposes, but it's not... I don't know, maybe

I'm just being overly cautious, but as a researcher, I'd really want to know where

the data came from and how it was collected.

PS15 later gathered information directly by "talking to [the original investigator] and learning

more information from that." However, because PS15 had to rely on "having them try to recall

and think back to what they did," he considered the information "limiting."

Even if participants worked with well-documented data, sometimes they still wanted

additional information. PP10 said:

PP10: You don't necessarily understand everything that was done in the study,

and so you do have to ask a lot of questions and make sure that you have a really

good understanding of what they did because otherwise it's very difficult to write

up the methods and exactly what [they] did. (…) It's difficult to write up all those

details unless you have a really good working understanding of them.

Others noted that "there wasn't enough information like I wanted" (PP13); "I just wanted to

know a little more about their samplings to understand how they did [it] and why they did [it]"

(PP15), and wanted to know "the logic behind their decision" (PS07). What PP10 and many

others discussed is consistent with the literature, which notes the inherently insufficient nature of

documentation and the limitations of knowledge transfer from the original data creator to

subsequent data re-users. In addition, due to the variance in participants' experiences and the

tacit knowledge needed to interpret data documentation, some may have needed more

information than others to reuse data.

3.3.2. Process of trust development: Provisional trust judgment

3.3.2.1. Changes in the data reusers' initial trust judgments

From the direct interaction with and investigation of the data, data reusers developed their trust. Ideally, the level of trust should remain as high as the initial trust or be increased by the strengths of the data from the direct interaction with the data so that the data reusers continue using the data. For example, 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 judgments became more negative when they started to use the data. PS06 explained working with the data made her have a realistic view of the data:

> PS06: It's like someone you're in a relationship with. We have a relationship with
> the data; we have high hopes and then we have… The honeymoon period is over,
> and then after a while you have a realistic view and relationship. (…) Knowing
> more about the data, working extensively, and having more experience can
> increase the level of trust but not always.

PS08 also described how her trust judgment became less positive:

> 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..."

Thus, provisional trust judgments often changed when the researchers actually used the data. This reflects the dynamic nature of trust, as the process of developing a final trust judgment also reflects (See section 3.3.3.).

Because provisional trust judgments were built on the reusers' independent judgments from interactions with the data rather than just on other people's opinions and perspectives, the participants' self-confidence was engaged. Participants showed their confidence by making judgments of the data; as PS09 said, "I trust myself," in an "intuitive sense" (PS08). Usually the confidence was accompanied by a good deal of past experiences using secondary data. For example, PP12 said, "I have a pretty good sense 'cause I've worked with a relatively consistent type of datasets so I have an idea about them." Similarly, PP01 "in general ha[d] a pretty clear sense" from experiences with "enough data sets to make judgment," and PP02 felt "very confident in my judgment on that."

3.3.2.2. Confirmation

Data reusers' provisional trust was developed through the process of *confirmation*. During confirmation, data reusers had both initial interactions with the data facilitated by acquiring them and in-depth exploration of the data. Trust was developed from both types of experiences with the data, and this was the process of confirming whether the data met the data reusers' trustworthiness expectations. Participants discussed various aspects of the trustworthiness of data as well as how they made their decision on trust.

Perceived preparedness of data

One aspect of trustworthy data discussed by the participants was their preparedness: how

well the data files and associated documentation were organized, how clean the data were, and how the data websites (when participants downloaded data from institutions, such as the CDC) were organized and presented. The preparedness of data that participants discussed during the confirmation process was mostly in relation to their feelings and impressions, which was a cognitive cue for trust. Their discussion of the data was at the presentation level (e.g., how information about data is presented to them), but the in-depth content of the data was not necessarily discussed in a logical fashion.

Some participants remembered the first time they discovered and viewed the data. PS09 recalled reservations after receiving a data package that consisted of data files and other information that was "not cleaned and organized at all…[I] was like, 'uh-oh.'" PS09 took this first impression of the data as a warning. In contrast, PS08 had a positive first impression of the organization of the files in a data package, which signaled their 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
> [the files] together. This is a general indication to me that the data is more likely
> to be trustworthy than not, as opposed to something that's kinda thrown together,
> hodge-podge.

This was a similar reaction to that reported by other participants when they first viewed publicly available data on websites. The appearance of the webpages and the organization of the information on those webpages can create both positive and negative biases. PP03 recalled her first experience with data available on a website as "very nicely done and very nicely laid out," and "it was very easy to find the information or the files that I needed." PS15 said, "in terms of

126

what they have on [the website] and how it's laid out," it was "very user-friendly." These interactions preceded a closer examination and downloading of data files or documentation; just by looking at the data's websites and forming impressions on how things were organized in the websites, some of the participants reported "a good feeling" (PH03) about the data, assumed their "trustworthiness" (PS08, PH03, PS15), and noted that they "definitely will use [the data] if [they have] right variables…if not, for my future project" (PS15).

This cognitive trust may not guarantee the real trustworthiness of data. What "very well done, very professional" means might depend on the criteria that participants have, and participants' level of trust may change as they interact with the data. However, it clearly indicates what participants expected of trustworthy data; they were well-prepared and carefully organized. The first impression about data, thus, provided a cue for their judgment. 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." From this, it is possible to infer that well-prepared data required some effort from those who manage and prepare data, which is associated with its trustworthiness.

Data validity

Participants asked questions, such as "are the data accurate?" (PP12) and "are the measures adequate?" (PS12). Data validity was one characteristic of trustworthy data, as it proved the data (variables and measures) were well-founded and accurately collected. Participants said checking on validity "made me feel confident and trust the data" (PP17). Data validity is an objective quality of data; therefore, the participants' trust was based on their rational judgments.

Although participants considered validity to be an important element in the development of trust, they also discussed difficulties in checking different types of validity on their own, like construct validity, by saying "in a way, you don't really know" (PP18). Thus, participants adopted a simple method to verify one aspect of validity on their own, "a sort of face validity thing" by "looking at the data to see whether or not they make intuitive sense" (PS08). The most common ways of verifying "face validity" were to run simple, basic, descriptive statistics, or checking the ranges of data through the descriptive statistics, whether "the ranges, the mins and maxes, are believable for these data" (PP07).

> PS08: Making sure that what you receive actually makes sense, and all of the responses are within the appropriate response set, checking any outliers, that kind of thing. (…) For example, doing basic descriptives; if something's not normally distributed, is that to be expected? For example, in alcohol data, you don't expect that drinking data's gonna be normally distributed. It's, almost, I would say 99.9% of the time never normally distributed. So, if I'm getting a data set that shows me a very nice and neat, evenly distributed normal distribution of drinking data, I'm gonna be suspicious. So, I think it's those kinds of things that I would look for.

Checking the distribution and ranges of data did not guarantee the validity of all data (e.g., validity of the measures or scales used in the study), but participants took this face validity as one indication of data validity as a whole and as an indication of it having been properly collected.

PP07: [I]f you see things, [that] there are a lot of unbelievable values, you start to

get the sense that maybe the whole data set is messed up, and it's not trustworthy.

I think of the actual values, and that at this point, I would believe how the data

were collected.

When participants needed to take a closer look at the data, they took some extra steps.

For instance, they searched for "trend[s] that we would expect based on what we know from

other national data" and required that "the initial frequencies of the variables of interest have to

make sense" (PP12) when the data were weighted to be nationally representative. If participants

did not find them, they suspected "the data [set] is not right, or it's not trustworthy" (PS19). One

looked for "basic relationships established in [her] field; are they there or are they panning out?"

(PS16). In a few cases, participants conducted a preliminary data analysis on the variables of

interest, and if the results were not "positively associated with the relationship quality, it would

then start to bring an element of distrust into the data" (PS15).

Documentation quality

Good quality documentation was the hallmark of trustworthy data. Participants discussed

different reasons that underlie their trust development based on good documentation.

Documentation not only provides the contextual information to understand data but also offers

information relevant to the scientific process and rigor in original studies; good documentation

reflects the original investigators' commitment to and concern for the data.

More information about data was more useful for participants' reuse and more detailed

documentation was regarded as good quality documentation for participants. Good

documentation creates more information on which to make a trust judgment. As PS14 noted, without documentation, participants cannot know "what's going on in the data"; only by knowing about data can they make a judgment. Reading the documentation was one way of assessing the study behind the data:

> PS08: You can tell from the documentation whether or not a research[er] was thorough and careful. If things are really well-organized and well-documented, then you can generally trust; I feel like you can better trust the data.

> PS15: If the [original investigators] don't really [have] anything else to send you, it's probably a red flag. Well, they might not actually really have all their ducks in a row, so to speak, in terms of really having the full documentation so that I would feel comfortable with this.

The participants believed that good documentation contained information that could prove the scientific rigor of the data, which was the participants' assessment of methods to create data as follows: the appropriateness of the study design to produce the data, the data collection mechanism, measurements, and even information about the data collectors if possible. Participants took "scientific rigor" or "process" as an important component of a trustworthy study, from which data were produced: "[H]aving scientific process and rigor" made the data "reliable and trustworthy" (PS12). Documentation was the only way to judge this scientific process and rigor (unless participants had received the information from a direct conversation with the original investigators): "There's a level of rigor that I attribute to that, that has been confirmed by what I've seen in the documentation" (PP01).

Documenting the research process and data production may not necessarily equal scientific rigor, as researchers could falsify the documentation. However, participants perceived that well-documented data were more likely to be found in scientifically rigorous studies because such documentation indicated the original investigators' commitment to and concern for the data and the study as a whole. As with the participants' perception of the preparedness of the data, good documentation created a sense of security among the participants.

> PS12: First of all, when you have stuff documented, that makes me feel more
>
> secure maybe because it means you wrote... You took the time... You are
>
> organized enough to put it on paper, and you can probably commit to it.

Participants often saw a connection between good documentation and the quality of the data: "[W]hen you spend times and put your efforts into something, it's less likely [to go] wrong" (PS17). PP03 noted that her trust increased based on documentation, acknowledging the effort behind it:

> PP03: I would say that once I actually get into it, I'm just even more impressed by
>
> the job that the individuals, scientists, or researchers have done in collecting the
>
> data... Cleaning up the data and putting them out there for others to use.

Transparency

Transparency was another characteristic of trustworthy data mentioned by the participants. For some participants, transparency was "the first word that comes to mind" (PS13) when they thought about trustworthy data and assessed trustworthiness. The characteristics of

transparency should apply both to the data documentation and to the ethics of the original investigators.

Transparency means that nothing about the data is hidden—including the data cleaning process, any changes made to the data, or limitations or issues with the data that the original investigators experienced. PP15 used data from researchers who were "very transparent about what they're doing," which him made "feel good" about the data. PS04 said that every change that the original investigators made to the data she used was well-documented, which increased her trust in the data:

> PS04: Everything was documented, like the changes that they made were all documented. So nothing was... Nothing was hidden. They knew the reason or logic behind it. Everything was purposeful. So, if they would say nothing about it or say like, "Oh, we just did it," then I think I would have lost trust. But they had a good reason to do that.

Recognizing the limitations of the study or data also gave participants "more [of a] sense of trust" (PP16). Several participants expressed their trust in the data when the original researchers were "being upfront" (PS05) and "honest" (PP17). One participant said "it seems more trustworthy" because the original investigators had not tried to "overstate" their results and "they are not claiming that this is the only way to answer these research questions" (PP16).

Consistency

Trustworthy data should be consistent. Participants discussed how internal consistency within data made them trust the data, which meant there was agreement between the

documentation and the actual data files and the same way of coding across all sources. Participants naturally sought consistency during their interaction with data files, reading documentation, and seeking information included in the data packages; they reported that inconsistency within the data could erode their trust in the data.

> PP01: I've never worked with a data set that I didn't feel like everything I was seeing was consistent. So, that questionnaire matches the codebook, and matches what I'm seeing in the actual data sets and code that they're providing. So I've always had that element of trust.

PS14 said the inconsistency in the use of scales and measures "really frustrates me," particularly in longitudinal studies. PP05 tells a story of starting with "pretty high" trust in some data, but this waned when she actually started to use them. She noted that "the biggest issues with trust sometimes come up when I first start using the data." She saw "a lot of inconsistency in the way the data were coded or things that just don't seem right" to her. Although she did not lose all trust in the data, "that makes me trust the data [a] little bit less," as she was not sure "what parts are wrong and what parts are correct." Similar to the participants' view on documentation and preparedness of data, a great deal of inconsistency between the data, documentation, coding process, and other factors was taken as a sign of the original investigators' inattention to detail and thus reduced their trust in the data.

Reliability

Reliability, the external consistency of data, which indicates whether the data would produce stable and consistent results across studies, was another characteristic of trustworthiness.

One way to check reliability was to run simple statistics of what the original investigator did and generate some graphs, assuming an analytical guideline had been published in the original study. Another way of checking reliability was by reading publications by other data reusers. While this process helped the participants know whether they had interpreted the data correctly, participants also saw it as a quick way to ensure that the data set was complete, could perform as expected, and produced the desired results.

Participants' checks on reliability were relevant to the predictability of data, as they could expect what the data would do and produce. Previous literature has often discussed predictability as a condition of developing trust. The reliability of data enhances predictability by ensuring valid outcomes of participants' results. PP16 explained why she was able to trust the data:

> PP16: It was relatively consistent with other kinds of studies that have looked at similar things. We didn't find anything that was totally out of the blue. It was consistent with [where] we began, consistent with what we expected.

PP13 found the data trustworthy because "the results [of the data] are consistent with what I have found with other data." The high predictability of data ensured the accuracy of the data analysis that participants would perform: "[the results] seemed consistent across the studies that I looked [at], that said [the data] is pretty trustworthy, (…) I believe it will do the same thing to me" (PP18).

Missing data

Missing data can decrease trust. Participants understood that there could "always be missing responses" (PS02) in data, whether from the survey respondents or by a mistake on the

part of the data collectors. However, too much missing data (responses or values) compromised the trustworthiness of the data. Although participants accepted some missing data that did not affect statistical analysis (e.g. "There shouldn't be too much missing, to have [a large] enough sample size of [a variable] (PP05)."), too much missing data made them question the accuracy of the data and the results of the original study. PP13 was working with data with many missing values at the time of the interview and questioned the accuracy of the data in terms of use and analysis.

> PP13: Well, there have been times where we had lower trust in the data in our examination of the raw frequencies. For example, a data [set] that we're analyzing right now from the [institution] matches to deaths from the National Death Index, and we have found that the number of deaths is too low. They seem to be missing about 18% of deaths. We're not sure if we can trust that data because that's a large portion of deaths to miss when the main reason we're using the data is to look at deaths, like educational status. (…) I wouldn't want that particular data set name to be published along with our concerns because we appreciate the [institution] working with us on that data, and we wouldn't want to publicly say that we thought there were problems with it. It's too early in the process. We're just talking back and forth and found [out] about the sources of the missing data and the accuracy of what they've sent us.

If a justification of missing data was provided, including how the original investigators had handled them in their analysis or for the reuse of prepared data (e.g., what PS02 called "how to treat, how to deal with them"), participants did not distrust the data. Transparency also

mattered with missing data, whether information about the missing data was provided through

documentation or through interaction with original investigators. Further, if the investigators

explained "what they [did] and how they [did]" it to make "valid interpretations of what the

results [were]" (PP15), it showed participants "the fact that they are capable [of making the data]

better, and that they know how to do that," which enhanced the participants' trust in the data.

Simple errors and mistakes

As with the participants' view on missing data—an inevitable outcome of human work—

participants accepted some human errors in data collection, the cleaning process, and

preparation, especially with "a large-scale data [set] with 100,000 subjects" and "thousands of

variables, [and] massive amounts of data" (PP16). PP16 thought some errors in data were "not

too surprising"; and for PP18, "it was impossible to expect all the data (…) to be perfect." Thus,

simple errors that were easily identifiable and fixable did not hinder its reuse and did not reduce

the participants' trust. PP02 worked with data that had a simple error, which did not change the

initial trust judgment concerning the data.

> PP02: So that particular variable was the only sort of mistake. But it was a simple
>
> fix, and it was actually easily detected once you sat and really thought about why
>
> your numbers look the way they look. So there was no change, as I looked
>
> through the data further. I felt the same way after.

However, participants interpreted simple errors and mistakes as errors in data

management, which may represent the original investigators' carelessness with the data. Too

many such mistakes can indicate a lack of rigor in overall research, scrutiny, and care for the

data.

> PS08: If you see a lot of careless errors, then it's much more difficult to trust the data because there may have been a lot of data entry errors. That means there weren't data checks.

Serious errors and mistakes

Participants sometimes discovered unexpected errors and mistakes in the data, which hindered their reuse. PP16 found errors in variable names, and after spending some time figuring them out on her own, she contacted the original investigators and said "they only came to realize an error when it was brought up to their attention." PS09 also identified errors in the data while checking their validity by replicating the original analysis, and found that "[the original investigators] basically [didn't] release that information."

Because serious errors in the data can directly influence the outcome of research, participants' trust in the data decreased when errors were detected. A few participants reported having found serious errors in the data. PP07 was working with data that a colleague had given her, but when she ran a model for analysis, the result made her go back to the data and re-examine it. Something similar happened to PS07; it influenced the results of the original study and likely his own. He was not sure that he wanted to continue.

> PP07: Once I got into it, there were so many issues in the data that were unbelievable. There were so many unbelievable numbers that it made me

concerned about the whole quality of the data set.

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. So, although

I had kind of gone through and looked at the sample size, a good funding source,

reputable investigators...

When PS07 found serious errors within data that "likely had affected the outcomes to a

certain extent," his "level of trust for the data went down," and he arrived at the following

conclusion:

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.

He reconsidered his data-selection process. Because he had already run a series of checks on

these data, the reputation of the original investigators, sample size, documentation, and other

checks, he wanted to add another criterion: "x amount of publications that had come from the

data."

3.4. Final trust judgments through problem solving

During the process of provisional trust judgment, data reusers' initial trust in data may

increase or decrease. Even though participants' trust may be decreased when they are faced with

data problems, such as issues with documentation and unexpected errors, there are still ways for trust to be restored.
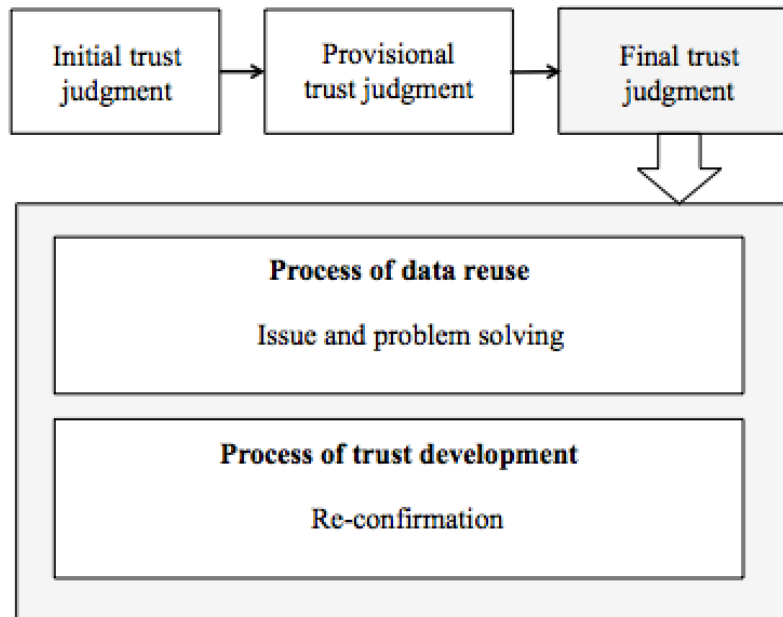
Figure 5. *Final trust judgment through problem solving*

3.4.1. Process of data reuse: Problem solving and helpful sources

Most participants sought answers solutions to problems and difficulties. They contacted sources that were connected to the data (e.g., original investigators, data-driven communities, and data repositories) or communities that were not directly relevant to data but trusted from previous research experiences (e.g., colleagues or research communities).

3.4.1.1. Original investigators (or their equivalents)

The most important players in recovering trust were the original investigators. Contacting the original investigators was the most common method of problem solving. These were the

original researchers who had created the data or their equivalent who were in charge of questions in cases when the data was produced by institutions. Several participants expressed a strong preference for access to the original investigators or their equivalent. They believed that direct communication with the original investigators offered "a behind-the-scenes look at the study" (PS08) and would illuminate "problems, issues, or simply questions" (PP02). PP04 explained that the greatest benefit of access to original investigators was as follows:

> PP04: They know what the challenges [that are associated with data] are and how to get around those challenges.

Those challenges are often difficult to explain and document, and participants said the additional help might not be possible without direct interaction with original investigators.

A close relationship with original investigators, particularly for the data from individual researchers, can be important as data reusers can ask questions informally. PS15 said, "you can be more candid with the individual researcher if you know them and have those more candid types of conversations" that can be more useful to understanding data issues. Communication styles that affected the types of conversations were also very different: "If I needed to write anyone at [national data name], it'd be very, like stock email, that it'd be very formal and it wouldn't be like, 'Hey, so and so...' It would be like, 'To whom it may concern.' (PS15)." He thought this different style and attitude of communication made the interaction with original investigators very different, which other people who were working with data without a relationship with the original investigators "won't really ever get" in terms of "one's comfort level coming in with the data and depth of information" (PS15).

Participants communicated with the original investigators over e-mail, in person, and by phone. For many of them, the interactions were helpful and increased their trust in the data. Original investigators "always responded very nicely" (PP06); they provided help by answering questions with a "fast turnaround" (PP09). This made a positive impression on the participants. In addition, because the original investigators "had a really in-depth understanding of the parent study" (PP02), interacting with them helped the participants surmount most issues they encountered. PP02 said the interaction with the original investigators was "immensely helpful" because it "provides a level of collaboration and mentoring that you don't see when you just get a dataset."

Unfortunately, interactions with the original investigators were not always productive. A few participants had difficulty contacting the original investigators. Although they attempted to contact them, the participants never received a response, it took too long to get a response, or they received only "partial answers," which was "very frustrating" (PS09). Sometimes participants found the interaction to be "tricky, as it just depends on if you can get the right person who has the knowledge [to answer the questions]" (PP12), particularly when they worked with large-scale institutional data. Usually the institutional data were associated with a contact person who may or may not have been directly involved in the data collection or original analysis. Although participants believed that the contact person would be an expert in the data and the right person to speak to, those negative experiences made participants wonder whether "all the staff were really busy or they may not exactly know" (PP12).

3.4.1.2. Data-driven community

A data-driven community was another good source for assistance. Participants interacted

with different types of data-driven communities, including a Listserv for specific data, workshops at conferences, webinars, and user group meetings. Because the original investigators sometimes ran the workshops and webinars, attending those sessions was another great way to contact them. The usefulness of the direct access to original investigators has already been discussed, and the workshops and webinars were helpful for the same reason. In addition, "there's a lot of training" in these workshops and webinars, which others also evaluated as "extremely helpful" (PP15).

Participants noted the presence of other researchers who had used the data in the data-driven communities and discussed the importance of their roles. PS04 said, "not just the original investigators, the project manager, and data manager but [also] the faculty members who worked directly with the data" were also at the workshops. Similarly, on the Listserv regarding the data that PP03 used, there was "a mix of university faculty, the university researchers, and federal researchers." Those researchers in this data-driven community were not necessarily the researchers in the same fields or disciplines of the participants but researchers with the same interests in the data. Neither PS04 nor PP03 recognized all of the people in the workshop or Listserv, and they did not have an established relationship with them, but they nonetheless expressed their appreciation for the help from these people in the community and the experiences the people shared with them. PP03 said, "it was more experienced researchers who answered the most…questions," which she found "super helpful" because "they'd know how to [solve the issue] but they'd also understand why I'm having this issue."

3.4.1.3. Colleagues of the participants

When contacting the original investigators was not possible, some of the participants

looked for alternative sources of assistance. Often, the easiest way to seek help was to ask accessible, reliable, and trustworthy colleagues. Participants used their colleagues as a source of data discovery and respected their opinions when selecting data: their own teams (or research groups) in the school, advisors, mentors, peers, and colleagues. PP04 attended a weekly team meeting where group members discussed their research progress, shared research ideas, and discussed challenges with data and other issues. Although the group meeting was not specifically about the data that PP04 was using and not everyone was working on the same paper, PP04 said "we'll discuss [the data], we'll bring it to the group and see what the group has to say because everyone has a different experience and expertise about things." For PP04, it was a "collaborative effort to solve issues" that was helpful when the members needed advice on dealing with errors and when one of the group had already worked with the data. By the same token, one participant tried to find "people around me who were more familiar, who could help me with understanding the data and the whole process" (PP14). Those who had worked with the data were usually their advisors, mentors, and colleagues, and participants felt "very lucky to have their expertise on the data set while I'm working on it" (PS06) when they were not able to obtain assistance from the original investigators. PS14 also shared:

> PS14: [My colleague] has worked with the data quite a bit. He has a lot of insider
>
> knowledge, which is very helpful. And he knows a few people who've worked
>
> with the dataset as well.

Some participants had colleagues who introduced them to someone who could help. PS13 had a colleague who said, "Oh, I have a friend who knows a bunch about it," and initiated an e-mail exchange. PS13 said, "it was much more informal" than the interaction with institutional staff

and was "extremely helpful."

3.4.1.4. Broader scholarly community

Participants also talked to other scholarly communities, which could include online forums, conferences, research meetings, and even their own team meetings at their institutions. The purpose of these communities was not to focus on data reuse experiences or problem solving. However, in communities whose members had the same interests and supported each other's research, people help each other with data issues and questions as much as they can. PP05 said, "It's pretty common in my field to go online and look for forums or discussion boards about [a] particular topic, which can be a data topic." PP08 also talked about getting help from a forum that was related to her research. According to PP08, the forum was "pretty active," and the members "share a lot of things about research, ask each other questions, answer the questions"; data-related questions and discussions were just one of the common themes in the forum. Because they already "talk about what they were using in terms of secondary analysis of existing data," it was not a "weird thing to do if I [brought] in a data question" (PP08).

Unlike a data-driven community, these scholarly communities tended to be those to which participants already belonged and in which they engaged on a regular basis. Because these communities did not exist specifically around the use of certain data or data types, finding solutions or answers for data problems might not be guaranteed, and the chance of a data expert being in the community might be lower than in a data-driven community. However, PP08's community, a forum of research centers, was "very active," and "there's a lot of talk about data in general." Although she had not used the forum for problem solving, she read the questions on the forum and responded to one.

The researchers that the participants interacted with in the scholarly communities were often interdisciplinary in nature. Participants from health and social work collaborated with researchers in their own disciplines, but some of them also engaged researchers from several other disciplines. One participant noted, "I don't usually work with people in my field" (PS07).

The researchers with whom the participants interacted were a mix of people whom they knew along with strangers; the participants knew most of the people from team meetings, research meetings, or conferences, but they generally would not know the respondents from online forums and other large groups personally. However, participants often trusted these communities. PS08 said, "These people are the ones that I chose, the people that I have created relationships with." The participants' trust in these scholarly communities can be different in terms of the usefulness of the community to acquire data information, but these communities were typically sources of casual data information.

3.4.1.5. Involving statisticians

Sometimes participants approached statisticians or statistical programmers during the process of problem solving. A few participants talked about the need for statistical consultations during their data reuse and analysis and complained about a lack of institutional support, mostly at small teaching universities. When participants had statistical support available, the statistician might have been assigned by the statistical services at the institutions as a team member for a project, or could have been someone who was just "doing [a] favor because there wasn't an appropriate person to consult with" (PP13). Those who had access to statisticians said they "talked with them if we had a problem or commiserated," and the statisticians could let them know "if the problem is really me or the data" (PP07). When the issue was "more like a

programming issue of why my merge[d data set] wasn't working," the participant was able to get help from the statisticians. When the issue was "really about the actual data, like if the variable didn't seem to be reflecting what it was supposed to be" (PP13), the participant approached the original investigators directly.

3.4.1.6. Data repository

Participants contacted data repositories for help if they had acquired the data from those repositories. Those participants made initial contact with the repositories rather than with the original investigators. Although the participants in this study perceived the repositories as a "neutral" place, often as "a warehouse" (PS09), and they thought the original investigators were primarily responsible for any issues with the data, participants still expressed appreciation for the repositories' help. PS09 contacted one data repository when he found that some of his results did not make sense. The first response he received from the repository was, "we're not really sure," which made PS09 think of the repository as just a warehouse. However, later, the staff later offered to look at the data for him and helped him to solve the problem.

> PS09: The people who are warehousing it were totally different people. And those
> people could sometimes be more helpful because they felt like that was their job.
> It's like, I was not talking to the PI but was [instead] talking to people [who] had
> any energy and expertise about the data.

3.4.2. Process of trust development: Final trust determination

3.4.2.1. Re-confirmation

This process of solving issues and problems influenced the participants' final trust of the data. This restoration (or decrease) of trust is known as *re-confirmation*. Interaction and communication with different parties were the keys to re-confirmation, as these encounters provided a chance for the participants' trust to be recovered through the solving of problems. Solving these problems deepened the participants' understanding of the data. Interactions with different parties were helpful when they were supported by deep knowledge, ethical and transparent attitudes, and shared experiences regarding data; together, these factors decided the participants' final trust during interactions.

Helpfulness supported by knowledge

While difficulty gaining access to the original investigators did not directly lower the participants' trust, it did make the recovery of trust impossible. There was a better chance of restoring the participants' lowered trust when the interaction answered a participant's question or solved a participant's problem. While easy access to and friendliness of the parties the participants contacted greatly influenced the participants' experiences with the data, what really mattered was the parties' knowledge and expertise. This was particularly important for the parties directly relevant to data, including the original investigators and the data repositories.

PP02 said that knowing that the original investigators "had a really in-depth understanding of the parent study" helped her establish trust in the original investigators and in the data. Whenever PP02 had approached the original investigators with questions, those

147

investigators had given clear answers and explanations about the data "without hesitance," which helped reducing PP02's suspicion surrounding these investigators. PS04 reported a similar experience:

> PS04: So I had detailed questions that I wanted to ask. (…) If [the original investigators] didn't have the exact answer to my questions, or they didn't know who was responsible for it, or who created that variable, or who decided to include that variable in the data, then I may [have been] more suspicious about the data rather than trusting it. (…) They answered every single question that I had. So, I think my trust level went even up, since I first started using that.

Further, the deep knowledge about the data possessed by the original investigators and their contact people also made participants think "[it was] (…) very well-managed data" (PS04). PP15's repeatedly positive experiences with the original investigators helped PP15 build strong relational trust with the investigators, as she noted: "I know the kind of help that I could easily get from then. They've always been there. I've listened to them, I've talked to them, and I feel good about their knowledge and their experience, which is part of my trust."

The participants had similar expectations for the data staff (curators) in the repositories. They expressed their trust in the professionalism of the repositories' staff members, stating that these individuals were "people that had any energy and expertise about the data" (PS09). PS10 said "[the repository] know[s] everything about data, they do their best to make sure [the data is] corrected, and they're very responsive. So that goes a way to continuously building trust." The same type of trust was reported in previous studies regarding data reusers' trust in repositories. In one study, reusers believed that repository staff were "well trained" with "expertise" and that

they "were the best possible people working" on the data (Yoon, 2014). Although the participants did not speak directly to the original investigators in the present study, the professionalism and expertise shown by the repository staff helped to rebuild the participants' trust in the data.

Transparent attitude

The impressions the participants had of the original investigators during their interactions were important factors in the recovery of the participants' trust. Transparency of data has already been discussed as an important element to enhance the participants' provisional trust; similarly, the transparent attitude of the original investigators was even more important when the participants communicated with the investigators about errors and mistakes found in the data. As already noted, the participants corrected simple errors by themselves. If they found serious errors and mistakes, some of them stopped using the data, while others were willing to give it one more chance by asking the original investigators to solve the problem. Depending on the original investigators' attitudes and demeanors toward the reported errors, this process could help recover the participants' trust if it had been lowered. Several participants discussed the importance of the original investigators' acknowledgment of the errors in the data, as well as their attitude when correcting the errors:

> PS12: Yes, they make an error, but I . . . The fact that they are willing to report that error, builds my trust in them because as it says, you know the proper thing to do.

PP01: When they make a mistake, they post it. They pull the dataset, they correct

it, [and] they fix it. There's no doubt in my mind that if they're aware of an error

or problem then they just ignore. And they're actually probably more rigorous

than the people that . . . in terms of acceptance of small errors, that the

[institutional] data sources that I've seen would print a correction or a retraction

on the data website much more rapidly than some other professionals that I've

worked with, so.

Several others reiterated the opinions of PS12 and PP01: "As long as they're honest about [their

mistakes]" (PP18). This goes back to the issues of transparency and ethics of the original

investigators, as discussed in the section regarding the process of provisional trust judgment

(confirmation).

Shared experiences

Finding shared experiences around the data was not only important for problem solving

but also for building the participants' trust in the data. Often, the participants could find people

from different communities who had reused the same data. These people included those from

data-driven communities, those from broader scholarly communities, as well as colleagues.

Sharing experiences with other data reusers helped the participants feel more secure about using

the data. As PP03 noted, "we are kind of in same situation." PP03 also noted that these other

reusers "[would] also understand why [he was] having this issue." The sense of shared

experience made PP03 accept the other researchers' suggestions and trust their solutions.

Conversely, negative shared experiences can decrease trust. PS04 noted that "knowing other

people who were closely working with the data" and "talking among [themselves]" was "really

reassuring [for her] experience with [the] data." This communication gave PS04 "confidence" about solving the issue and continuing to use the data.

3.4.2.2. Final trust confirmed during data analysis

After the participants took all the necessary steps to understand and investigate the data, they moved on "to run the actual models to see what [was] happening" (PP05). To do that, some participants extracted key variables for their research or recoded them "in a way that [made] sense to [them]" (PS05). Depending on the research questions, the participants merged or combined multiple data sets. This process happened when they read the documentation, interacted with the original investigators, and checked the data for validity and reliability.

While it was possible to run into problems during data analysis, most of the participants said that their trust did not change at this stage because "[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). Some participants said that they never lost trust in the data during the analysis, whether "a small independent data set or [a] national data set" (PS15):

> PS01: I mean, not at the analysis stage. Obviously you still form your opinion when you're reading the literature and the documentation about the data, the documentation on sampling design, documentation on measurements, and this and that, but not while I'm handling the data. I make my judgment before that.

PS02 completed her trust judgment on the data before analysis and nothing "change[d] [her] opinions about the data." She continued:

PS02: I only question my ability to understand the data. I don't question the data.

If something is going on, it's not the data, it is something that you don't

understand. The problem is not the data; the problem is you.

The participants who were confident that they had made the correct judgment in terms of the data before analyzing it did not find any fault with the data.

Chapter 6. Discussion and Conclusion

In this chapter, I will first provide a brief summary of the findings. Then, I will discuss the importance and nature of trust in the context of data reuse, and present data reusers' trust from theoretical perspectives by mapping the trust attributes that have appeared in this study along with attributes defined by previous studies with implications for data curation. I will conclude this chapter with limitations and contributions of the present study. Venues for future research will follow, as the limitations of the study suggest directions for future work.

1. Summary of findings

The initial aim of this study was to understand how data reusers' experiences with reusing data relate to their development of trust. 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 their trust can also be formed, lost, declined, and recovered during their data reuse experiences. These various stages reflect the dynamic nature of trust. The participants also identified and discussed various trust attributes that influenced their formation of trust in data they explored and used.

The motivation to utilize existing data was the starting point of data reuse. Data reusers from the fields of public health and social work had different motivations, some of which were specific to their respective domains. These motivations included the potential and value of existing data, the cost-effectiveness of data reuse, the availability of large samples, and education

and training purposes. Often, these motivations were often related to the studies' data selection criteria.

When my interviewees needed to reuse data, they started looking for available sources (the process of data discovery) and pre-screened any data they found based on its relevancy and usability (the process of initial selection). These processes were relevant to their initial trust judgments, which were developed even before they had any direct interactions or experiences with the data. Initial trust may not guarantee full and final trust in the data, but it is nevertheless influential for data reusers when they are deciding whether they want to continue to the next stages of the reuse process (acquiring, investigating, and understanding the data for analysis). This study's participants developed their initial trust judgments through the processes of prediction, attribution, transference, and bonding. Because they had not actually used the data at this point, they relied on various trust attributes in forming their initial trust judgments. These attributes included their past positive experiences in cases when they went back to the same data, some observable evidence for judging the data, other people's assessments of the data, and their personal relations with the original investigators. The data reusers established this trust rationally, emotionally, and cognitively at the individual level (e.g., they were influenced directly by a researcher or by the original investigator), the institutional level (e.g., they were influenced by the institutions that produced the data or to which the original investigators belonged), and at the societal level (e.g., they were influenced by the academic culture of the data). A low level of initial trust did not necessarily prevent further investigation of the data, as some reusers still wanted to investigate the data themselves despite their suspicions. Still, initial trust did play an important role in determining data reuse, as this trust led the reusers to the next stage: acquiring, investigating, and understanding the data.

The data reusers' experiences of acquiring, understanding, and investigating the data varied depending on the data package they received, the condition of the files and documentation for that data package, and the nature of any interactions they had with the original investigators (if they had had any interactions at all). The process of investigating and understanding differed depending on the individual practices of the data reusers. In general, the data reusers reported reading the documentation, reviewing the publications associated with the original or reused data, interacting with the original investigators if the documentation was insufficient or if there was no documentation, and checking various aspects of the data. As the data reusers directly examined and investigated the data, they discussed the multiple trust attributes involved in the development of their trust judgments. For them, these attributes represented the characteristics that determine the trusted nature of data and influenced the data reusers' judgment of the data. Depending on the degree of trust violation or satisfaction, the level of the data reusers' trust could be decreased or increased.

Provisional trust could be the same as participants' final trust judgment if they found the data trustworthy enough or if there were no serious trust violations during the process of investigating and understanding the data. The data reusers had to decide whether they wanted to continue using the data when there was a trust violation or any other problem with the data. The process of searching for solutions to issues and for explanations of trust violations was part of the process of developing (or confirming) their final trust judgment in the data. The keys to this process were the problem-solving steps that were supported by the knowledge and expertise of the different parties with whom the data reusers tried to interact, including the original investigators, staff members of the data repositories, the data-driven communities, the reusers' colleagues and other researchers, and the broader scholarly communities to which the data

reusers already belonged. The data reusers' evaluations of their interactions with data-relevant parties were influential in the further development of their trust of the data. The transparent attitude of the original investigators toward any issues or problems reported was particularly important to the reusers' trust. Learning about other data reusers' experiences was also very influential because other researchers' positive experiences provided clues for solving issues and also offered some emotional support for the reusers to keep using the data.

2. Importance and nature of trust in data reuse

2.1. Data reuse and the relevant circumstances of trust

At the beginning of this paper, I argued that trust is and should be considered important in the context of data reuse by discussing conditions in which trust is more relevant. This study confirms the current situations in which data reusers meet these pre-conditions for questioning trust in data. Previous studies have argued that the question of trust becomes relevant when individuals meet the following pre-conditions: users' need for using information, lack of standards for ensuring information quality (uncertainty), and potential harm from using poor information (risk) (Kelton et al., 2008). The current situations resemble the circumstances of trust to be considered, which include the following: data reusers' clear need to reuse data, the lack of standards for ensuring data quality, and data reusers' awareness of the potential harm that could come from using inappropriate data for research.

First, the study participants discussed different motivations for reusing data. They all had clear reasons for reusing data, and each one understood the advantages of reusing data, such as cost-effectiveness and the potential of secondary data. Sometimes the need for reusing data is situational, as it is necessitated by the difficulties of collecting specific types of data with specific

samples (e.g., national samples). In that case, reusing data is not a choice but a necessity; it is the only way to answer certain research questions or to experience a specific type of data. The need for reusing data in these situations shows the dependence of reusers on secondary data.

The participants also had to deal with uncertainty regarding the data. Previous research has implicated the relationship between uncertainty and trust in various contexts, particularly regarding the role trust plays in users' decisions when they are made under conditions of uncertainty (e.g., Doney & Cannon, 1997; Gambetta, 1988; Luhmann, 1979). Uncertainty of data is often associated with (1) a large volume of generated data and (2) the lack of standards for defining quality data before the data are distributed. Some disciplines have practices to control data quality before preserving the data in trusted repositories and distributing the data (Adelman et al., 2010; Stockhause, Höck, Toussaint, & Lautenschlager, 2012); however, such practices have not been fully implemented as a normal practice in the social sciences, although related discussions in data journals and the conducting of data peer reviews have recently been emerging. The lack of systematic control of the data makes the assessment of the data the responsibility of the researchers conducting a secondary analysis, i.e., this study participants. Data reusers rely on the indirect assessment of data through the publications that have produced or reused these data. However, this indirect assessment of data does not always guarantee data quality. This trend was demonstrated in the present study, as the participants' trust judgments changed throughout the development of their initial and provisional trust. This alteration was caused by the potential differences between what was being produced and used in the original research and what was available for the data reusers in specific forms and formats.

The potential harms or risks were another element of trust to be considered. The participants were aware of the potential dangers (or risks) of using poor or inappropriate data

because they understood the importance of the role data play in research. The participants in this study were empirical researchers; thus they relied on data to "provide evidence" (PS04, PS12) for their claims. As empirical researchers, the participants understood the consequences of using low quality data; as PP15 noted, "garbage in and garbage out." This study demonstrates the relationship between risk and trust, in that a sense of trust encourages risk-taking behaviors, just as past research on this topic did, (e.g., Giddens, 1990; Mayer et al., 1995; Sheppard & Sherman, 1998).

While the results from this study demonstrate how and why the concept of trust is relevant to data reuse, the real question regarding trust among data reusers remains: How can the these preconditions of trust, uncertainty, and risk can be changed to increase the level of trust toward data? In other words, how can trust be fully established, thereby removing the need for preconditions of trust?

2.2. The dynamic nature of trust during trust development

Few studies have directly examined the dynamics of trust in data. Many studies that have examined this issue have discussed the nature of trust as a static factor rather than a dynamic one. In contrast to these previous studies, this study reveals the social and dynamic nature of trust. Trust can form, disappear, decline, be lost, and be recovered. Different decisions regarding trust are involved in different types of relationships. Trust in the context of reusing data engages not just the data but also various social parties, including the people involved in data creation and collection, institutions, and communities that have used and shared experiences with the data. This type of trust involves diverse trust attributes and patterns that arise from the decisions (see section 3. for diverse trust attributes).

158

A trust relationship begins with trust formation (Mayer et al., 1995; McKnight et al., 1998). This study refers to this type of formation as "initial trust development." According to past studies, the key process of trust formation is how the trusting entity (the trustor) can infer the trustworthiness of the entity being trusted (the trustee). During initial trust development, data reusers usually infer the trustworthiness of data from the characteristics of relevant parties around the data through the processes of prediction, attribution, transference, and bonding. This step is carried out before building a direct relationship with the data, and these processes are performed until the data reusers perceive the trustee as sufficiently competent and satisfactory. At that point, the data reusers are more likely to consider the trustee as trustworthy and will thus increase their trust of the data.

Various conditions can further foster trust; these conditions may also allow trust breaches and violations. While the development of initial trust leads data reusers to have positive expectations for the data, which encourages them to take the risk of spending time investigating and using data, there is also a possibility of trust violations occurring in the trust relationship with data. In the present study, a number of participants gained trust during the development of provisional trust, which involves directly building a trust relationship with the data following thorough investigations. However, some participants' trust declined after one or more violations had occurred. As Fulmer and Gelfand (2013) explained with their concepts of vigilance, sustained attention, and alertness, when data reusers are vigilant toward violations, they may be more likely to notice violations and to lower their trust. When data reusers identify violations, they must then decide how much they should lower their trust in the data. While this study did not quantify changes in the levels of data reusers' trust, the study participants did show different attitudes regarding various types of violations. For example, some participants were willing to

159

accept small errors in data (e.g., easily noticeable and fixable errors) but not serious errors (e.g., errors that are not well-explained by the original investigators, that are not easily fixed, or that directly impact the data analysis). Thus, the awareness of a violation does not always prompt responding actions: some participants said such violations immediately reduced their trust, while others were still willing to withhold decisions regarding their continued use of the data. These results suggest that data reusers may not change their trust attitudes or lower their trust based on a single violation. This conclusion also suggests that not all trust attributes carry the same weight in changing data reusers' trust levels.

Final trust judgment is dependent on trust restoration when trust has been decreased, which is relevant to data reusers' responses to violations and to their final decisions regarding the data. Proper justification of the violation is key to the process of trust restoration. Previous research regarding trust in human relationships explained this process using the concept of "idiosyncratic credits" and the latitude of acceptance. The trustor offers the chance for the restoration of trust based on the accumulation of positively disposed impressions residing in the perceptions of relevant others, and the trustor also has a range of accepted and tolerated positions (Hollander, 1958, p. 120; Hovland, Harvey, & Sherif, 1957; Sherif & Hovland, 1961). Similarly, data reusers rely on various relevant parties' opinions and experiences when solving problems and violations. The restoration of trust is generally predicted when the trustor believes the violation to be unintentional (Tomlinson & Mayer, 2009). Data reusers consider the intention behind violations (e.g., data errors) as well as the transparency of the trustee (including the data and the original investigators) to be important for trust restoration. These reusers want to make sure any violation that has occurred was situational and unintentional.

As the previous discussion suggests, data reusers make trust decisions across three trust

phases—initial trust development, provisional trust development, and final trust development. Levels of trust can fluctuate at divergent rates in different phases. Different types of interactions with relevant parties concerning the data and different trust relationships form dynamic patterns of trust development. While these dynamics may be innate characteristics of trust, a less dynamic pattern might be preferred in the reuse of data, as the dynamics are often related to various trust breaches and violations (although there are changes for restoring trust after the breaches). Well-curated data could help easing the process of trust judgment by eliminating the factors that cause trust breaches and violations. Details about data reuser-defined trust attributes for data curation are presented in the following sections.

3. Trust attributes identified during the trust development

This study identified the different trust attributes as established by the processes of participants' trust development. Several of the attributes established in this study were conceptually similar to the trust attributes defined by previous studies. Some new attributes were also established. Using the trust attributes presented in Chapter 5 (sub-trust attributes in Table 8), I re-categorized them into higher levels (higher level trust attributes in Table 8) so that they are mapped along with the trust attributes from previous studies (Table 8). Some trust attributes played a role both in forming and recovering the participants' trust during the different processes of trust judgment (e.g., transparency during the provisional trust development and the final trust confirmation); categorizing trust attributes in a higher level presents an easier view of overall trust attributes without overlapping concepts.

While previous studies only viewed and discussed trust attributes as antecedents, the dynamic nature of trust that appeared in this study (see section 2.2. of this chapter) suggests the

need for understanding trust attributes using broader perspectives. Antecedent attributes helped to develop initial trust while confirming and supporting the types of trust attributes that were used to develop further trust. These confirming attributes helped determine the final trust judgment from the reusers' own assessment of the data, and the supporting attributes reinforced the reusers' trust. The types of trust attributes are not mutually exclusive, as several attributes are relevant to multiple types: for example, both commitment and transparency are antecedent attributes that can confirm reusers' final trust by contributing to the data properties (e.g., documentation).

Overall, the trust attributes defined by data reusers reflect the social nature of data. Data reusers developed their trust from multiple aspects of associated entities rather than relying on one entity through calculus-based, affect-based, and cognitive types of trust. These trust attributes also reside in various levels, including the object (data), individual, institutional, and societal levels. At the object level, the intrinsic characteristics of data properties within the data themselves, such as validity, rigor, and reliability, are important attributes with the information about data, such as comprehensiveness. Individuals and institutions that are relevant to the data are another important part of users' trust, and users' trust in each can supplement each other. Trust is also involved with society in general or can also be specific to the communities relevant to data reusers

Table 8. *Trust attributes in data reuse*

Higher level trust attributes	Sub-attributes of trust	Trust development	Types of attributes	Trust antecedents from previous literature	
				Trust in human relationships	Trust in digital information
Ability	Competence of OIs; Helpfulness from expertise	Initial trust; Final trust	Antecedent; Supporting	Ability	
Ethics and integrity		Initial trust	Antecedent	Benevolence; Integrity	Objectivity
Predictability	(Based on previous experiences)	Initial trust	Antecedent	Predictability	
Commitment	Documentation; Preparedness; Consistency; Mistakes	Provisional trust	Confirming; Supporting	Identification (in an organizational context)	
Transparency	Transparency of data & original investigators	Provisional trust; Final trust	Confirming; supporting	Transparency (in an organizational context)	
Social acknowledgment	Reputation; Recommendation; Publications	Initial trust	Antecedent		Source of authority
Rapport		Initial trust	Antecedent		
Community reliance	Shared experiences; Helpfulness	Initial trust; Final trust	Antecedent; Supporting		
Preparedness		Provisional trust	Confirming		Accuracy
Comprehensiveness	Documentation	Provisional trust	Confirming		Completeness
Scientific rigor	Validity	Provisional trust	Confirming		Validity

163

Ability

Previous studies of trust in human relationships have suggested that the ability of a trustee is one of the common trust attributes (e.g., Deutsch, 1960; Mayer et al., 1995; Sitkin & Roth, 1993). While competence and expertise are synonyms for this ability proposed by previous studies, the present study uses the term "ability" as suggested by Mayer et al. (1995) because this term highlights the context-specific nature of the definition. In this study, the ability of the parties that were directly associated with the data—including the original investigators (both individual researchers and institutions) and data repositories—is an important trust attribute. For the original investigators, ability can be described as research competence and expertise in the related domain. Ability often indicates expertise and skills in methodology; in this study, it refers more specifically to skills in quantitative methodology. For data repositories, ability is related to the repository staff's professionalism, knowledge, and expertise. Ability contributes to the development of initial trust in data through the processes of attribution and transference by checking the communities of practice to which the original investigators belong, as well as verifying the reputations of the original investigators and repositories. Ability also helps confirm data reusers' final trust judgment during the process of re-confirmation through the direct interactions with the original investigators and data repositories.

Ethics and integrity

The ethics of both the original study (that produce data) and data producers (the original investigators) are another trust attribute. Several researchers have argued that positive intentions and motivations of the trustee impact the trustor's trust judgment as well as the trustee's integrity, which implies a strong sense of justice (e.g., Deutsch, 1960; Giffin, 1967; Lewicki,

McAllister, & Bies, 1998; Mayer et al., 1995; Sheppard & Sherman, 1998). In the context of information, positive intentions of a trustee are understood as the presentation of information that is free from deception or distortion (Kelton et al., 2008; Lucassen & Schraagen, 2011). Possibly due to the roles the original investigators played in collecting the data, the participants of this study considered the ethical characteristics of the data to be the same as those of the original investigators. It was important for the participants to know the intention behind the data creation. They needed to determine whether the data had been collected for the public good (i.e., for science) or for someone's private profit. They also needed to identify the data collection process to decide whether the data had been created in an honest manner, which lacked deception and distortion. The moral motivations of the original investigators were extremely important for supporting (and ideally guaranteeing) the objectivity, integrity, and ethical quality of the data. Thus, the ethics and integrity of the original investigators were important considerations for the participants during their initial trust judgment attribution.

Predictability

Kelton et al. (2008) proposed predictability as one attribute of trust. Other researchers have used different terms, such as reliability (Gidden, 1990) or consistency (Butler, 1991), to refer to predictability. Predictability is associated with data reusers' expectations of the data's reliability and consistency as determined by past experiences using data and cumulative experiences working with the same original investigators or data repositories. Because the participants used their own experiences to assess predictability, they were more certain about the data and about their own judgment of trust. Predictability is an important part of developing initial trust, but it might not be applicable when data reusers are working with new data.

Commitment

The trustor's understanding of the trustee's interests based on shared values and commitment is seen as an important trust attribute in the context of stakeholders' trust in an organization (Lewicki & Bunker, 1996; Prison & Malhotra, 2011). A similar type of trust attribute appeared in this study; this was known as the attribute of commitment. The participants of this study cared about the relevant parties' commitment to the data; this commitment level influenced the development or enhancement of each participant's trust. The participants described different ways of evaluating the commitment to the data. Most of these methods were related to preparation activities for data sharing. Demonstrations of both perceptions of commitment and evidence of commitment helped the parties develop the reusers' trust in the data. The participants were able to judge each party's commitment to the data from the participants' initial feelings and impressions of the data, which could be associated with each party's preparedness, the organization of information in data packages, and/or the presentation of information on data websites. The participants also found evidence of each party's commitment to the data during the development of provisional trust. Such evidence was discovered through well-prepared documentation, consistency of documentation and data packages, and the presence of only a few small errors. Although this study's participants usually considered the original investigators to be actors who showed commitment to their data, some participants also consulted data managers when citing the need for commitment. Thus, in a data reuse context, the expectation for commitment was not just limited to the original investigators but to anyone involved in data preparation or curatorial activities.

Transparency

The previous literature described transparency mostly in an organizational context. More specifically, this literature examined whether organizations were willing to share trust-related information (Prison & Malhotra, 2011; Mishra, 1996; Tschannen-Moran, 2001). Transparency also matters in a data reuse context; when data reusers find that information about data is transparently documented or if they perceive transparency in the original investigators' attitudes regarding their data—particularly in respect to data limitations and errors—they are more likely to trust the data. Transparency is one trust attribute relevant during provisional trust development (which occurs during the process of confirmation); it is also pertinent during the final trust development (which takes place during the process of re-cofirmation) by helping to recover the trust diminished in the previous stages of data reuse or by increasing the level of trust. While the importance of transparency has been recognized (e.g., by the Committee on Science, Engineering, and Public Policy [U.S.], 2009), Coates (2014) argued that transparency has not yet become a standard practice. Coates (2014) maintains that there is a need for an increase in transparency; such an increase could occur if greater support for data management practices were provided as well as other research activities (e.g., data citation).

Social acknowledgement

Previous studies regarding trust in information have already noted that trust can stem from social acknowledgement. Such acknowledgement may include the reputation of data and recommendations for using data. The trustor is more likely to trust if others also trust the trustee (Kelton et al., 2008). As a socially constructed concept, trust can be transferred in the context of data. Reputations and recommendations for data as well as for the original investigators were

important for reusers to develop their initial trust in the data (transference). The participants

accepted the data and original investigators when recommendations came from accepted or

trustworthy sources, such as advisors or senior researchers, colleagues, and research

communities acknowledged by the reusers. Peer review and evaluation are the formal form of

social acknowledgment regarding data; these elements helped reusers develop trust during the

process of attribution. The data reusers also clearly relied on indirect peer assessment of the data,

such as peer reviews of publications that produced or reused the data; this reliance was possibly

due to the lack of formal and direct assessment of the data.

Rapport

Rapport, which results from a close relationship, is strongly rooted in affective

interpersonal trust. Rapport was one of the newly added trust attributes in this study. Because

affective interpersonal trust is influential in most human relationships, rapport becomes a trust

attribute when data reusers work with researchers with whom they are already in a close

relationship. A close relationship is helpful because it allows the reuser to know more about the

original investigators' ability, ethics, and integrity; this knowledge helps reusers trust the original

investigators and their data. While rapport is an influential antecedent for trust development, it

does not guarantee the reusers' final trust in the data, as noted by a few participants.

Community reliance

A shared experience with the data is useful for reusers, as it gives them a feeling of

reliance based on their group experiences. To become a trust attribute, a shared experience with

data should be positive, helpful, and acceptable to data reusers. The positive experiences of the

people around the data reusers work as an antecedent attribute, as it helps the reusers develop

initial trust in the data. The feeling of reliance gained from shared experiences is more important for confirming reusers' final trust judgment than it is for enhancing the reusers' confidence in the reusing of the data.

Preparedness

Accuracy or correctness was a trust attribute identified by previous studies in the context of digital information; accurate information is error-free (Lucassen & Schraagen, 2011; Kelton et al., 2008). In a data reuse context, preparedness is a trust attribute that is similar to accuracy or correctness. Preparedness in terms of accuracy or correctness is different from validity, as it is related to the readiness of the data being reused and not to the methodological soundness of the data. Datasets may contain errors and mistakes caused by carelessness in the process of managing data. Some common issues identified by the participants of the present study included inconsistency of terms used in documentation and data packages and disorganized documentation and information. Well-prepared data without those issues helped build the participants' trust, and it was also related to the trust attribute of commitment (as participants considered preparedness to be evidence of commitment to the data).

Comprehensiveness

Users of digital information are most likely to trust that information when it is complete (Lucassen & Schraagen, 2011). In the context of data reuse, reusers are concerned with the comprehensiveness of the information provided to them about the data. The importance of the comprehensiveness of this information is significant, as data reusers rely on this information to understand and assess that data for validity, reliability, and trustworthiness. Documentation is usually evidence of comprehensiveness, although it is also possible to obtain information directly

169

from the original investigators. The participants in this study had both positive and negative experiences obtaining information from the original investigators; the delivery of that information was sometimes limited by unreliable memories, difficulties in interaction, and accuracy levels of the information.

Scientific rigor

Previous studies regarding trust in digital information consider validity to be the use of accepted practices and sound methods for creating information (Kelton et al., 2008). Validity can be directly applied to data reuse; reusers consider validity an acceptable way to create data, to develop a study design, and to take measurements. The participants of this study often used the term "scientific rigor" in a similar sense. Validity is a strong confirming attribute of trust, and the participants of this study checked the face validity of the data (e.g., the raw frequencies of simple variables) as one way of ensuring different types of validity. Documentation was also used to verify the validity of the original study because the documentation contained information relevant to the scientific rigor of the study, including information about methods, measurement, study design, and procedures of collecting data.

4. Implications for data curation

As discussed above, the changes in the data reusers' trust judgments during their discovery of and experiences using the data demonstrated the dynamic nature of trust. Trust can form, disappear, decline, become lost, and be recovered, depending on the condition of the data and the different trust attributes that are engaged in the process of trust judgment. The trust attributes that appeared in this study suggest various implications for data curation, including the following: 1) personal relations with the data creators, other data reusers, and repository staff

have an impact on reusers' trust judgments due to the lack of systematic support in data reuse practices, which presents the need to develop a method to support data reuse through formalized systems; 2) data management activities influence the likelihood of data reuse as they directly influence reusers' trust judgments; 3) initial data management work conducted by the original investigators plays a significant role; 4) the intrinsic properties of the data need to be integrated into the process of data curation, and a range of different stakeholders should be involved in data curation activities; and 5) user communities need to be involved in data curation activities.

4.1. Overcoming data reusers' reliance on personal relations and characteristics

Several data reuser-defined trust attributes, including ability, ethics and integrity, and rapport, suggested that the personal relationships and characteristics of the original investigators were a big part of the data reusers' trust judgment. This was especially true when the participants acquired data directly from the individual researchers who had produced them. A close relationship with the original investigators and a recommendation from people who were close to and familiar with these investigators influenced the building of the participants' initial trust in the data. Participants also often transferred their positive feelings regarding the original investigators to their positive attitudes toward the data. This reliance on personal relationships and characteristics for trust judgment reflects the lack of a formalized system for data sharing and reuse practices.

Several scholars have argued that trust is embedded in the impersonal structure of high-trust societies, which means that trust is carried out by formalized institutions and professional organizations and systems on behalf of individuals' interpersonal relations (Fukuyama, 1996; Gidden, 1990). On the contrary, a lack of systems or formalized institutions can cause people to

171

rely too heavily on individuals. This is the case for the current situation of many data reusers. Although the benefits and growing needs of data reuse have been acknowledged, practices of data sharing and reuse are not entirely systemic, nor are these practices fully supported by institutions. While the participants of this study expressed their interests and needs for diverse data from various data producers, they often faced difficulties in discovering available datasets, particularly from individual researchers and research teams. Further, the forms of information distributed to the study's participants were not standardized. The lack of control in the data for reuse was related to the level of uncertainty that data reusers have to deal with (as discussed in section 2.1. of this chapter), and personal relations were often the easiest and the most assuring way for the reusers to judge trust.

As noted earlier, data are a social product; several different parties are engaged in data production, evaluation, and use. The dynamic social relations and interactions surrounding data are often locally embedded. These relations and interactions may even be the reasons that data reusers' development of trust is a dynamic process; however, the question remains as to how those social relations and interactions can be effectively supported through formalized institutions and systems.

4.2. Importance of data management for users' trust

The study results and user-defined trust attributes indicated varying degrees of data reusers' awareness of data management. The participants of this study showed different levels of awareness of data management. Whether or not they referred to this work as "data management," most of the participants showed at least some awareness of data management activities, including the data preparation work performed by the original investigators or relevant project personnel.

A few participants expressed their awareness of data management by saying, "it's on the data management part" (PP02) or, "it's really well-managed data" (PS04). Although it was not clear whether or not the participants were aware of all the data curation activities, such as preservation, it was obvious that the participants were aware of some data management work during their reuse experiences.

Data management activities directly impacted the participants' trust in the data they worked with. Several trust attributes were used as evidence of "well-managed data," such as the comprehensiveness of the information provided and preparedness of the data, including proper documentation, good file organization, the use of labeling, and the absence of errors. Other trust attributes, such as commitment and transparency, were also part of the data reusers' concerns for data management, as these attributes indicate the manner in which management work must be conducted.

While it is interesting that data reusers' trust judgments are directly relevant to the work of data management, some attributes are not entirely new considerations in data management. For instance, practices for documentation have been well discussed and supported by recommended practices of data management and curation (e.g., ICPSR, 2009; UK Data Archive, 2011; UC3, 2011), as well as tools developed to support data management (e.g., DMP tools).

However, some aspects of these attributes can be subject to interpretation, and further understanding of the reusers' perspectives is thus required. For instance, comprehensiveness of information can mean different things to different data reusers, as each of these individuals have distinct tacit knowledge and research experiences. While there are different requirements or recommendations for the description levels and metadata from funding agencies and data

repositories, it is important to find an agreed-upon level of comprehensiveness from the users' perspectives. At the same time, a balance between the users' expectations and the original investigators' burdens must be achieved.

Other attributes, such as commitment and transparency, may not always be apparent to data reusers even though they take them into account for developing their trust. While the participants of this study often relied on indirect indicators of commitment and transparency, such as preparedness of data and further interaction with the original investigators during the reuse, more visibility would be helpful for developing the data reusers' trust faster and more easily.

4.3. Roles of original investigators in data management

While it is apparent that data management work influences users' trust development, the role of the original investigators in data management was particularly important for the study participants. The majority of the participants of this study discussed their experiences acquiring data directly from the original investigators (both individuals and institutions). Those experiences were different when using data from institutional repositories: the parties that perform intermediary roles and have expertise in data curation.

In circumstances where data sharing and reuse are conducted among individuals without intermediary parties, the original investigators' understanding and experiences of data management were significant for ensuring the reusers' trust in the data because the data that participants worked with were prepared and managed directly by the original investigators. While it was difficult for me to know the actual depth and thoroughness of the data management work conducted for the data used and discussed by the participants in this study, the participants

experienced different levels of data management work done by the various original investigators.

The previous studies have already argued for the importance of the roles of original investigators (data producers) and defined their responsibilities in data curation (e.g., Humphrey et al., 2000; Lyon, 2007; National Science Board, 2005). The activities that should be performed by the original investigators are also the starting point of all curation activities in the DCC curation lifecycle model through the phase of "Conceptualise," although the model does not visibly specify the engagement of the original investigators. In the data curation lifecycle, the roles of original investigators are essential to ensure further curation activities performed by data curators, but this study adds the additional importance of their roles as their work directly impacted the users' trust judgment. Educating the original investigators regarding the needs of data management and engaging in the early stages of data creation to ensure proper data management are not just further value-adding activities by data curators; they also support trusted data sharing and reuse among individual researchers.

4.4. Curating intrinsic properties of data

The results of this study indicated that the intrinsic properties of data, such as the validity, reliability, and scientific rigor of data, are an essential part of data reusers' trust judgments. It was also indicated that data reusers assess the intrinsic properties of data through thorough investigations. Data reusers' trust can be confirmed using the intrinsic properties of data, and higher levels of trust determined by other trust attributes during initial trust development cannot guarantee reusers' trust in data if trust attributes relevant to these intrinsic properties (e.g., scientific rigor) are not present.

Despite the significance of intrinsic properties, previous studies have questioned whether the intrinsic properties of data can be or should be curated in a curation context. While the intrinsic properties of data are the responsibility of the original investigators (i.e., the data producers), the issue of curating these properties concerns appraisal and selection in the curation lifecycle model. Whether or not data are appraised and selected is relevant not only to long-term preservation and curation of the data but also to the validity and reliability of the content (or research) itself. DCC's data curation lifecycle includes the selection and appraisal of data for curation; however, selection and appraisal are usually linked to the repository or institution's collection development policy and legal requirements. Selection and appraisal are also associated with the potential for long-term curation and preservation of the data, which do not necessarily consider the intrinsic properties of the data. McDonough (2012) argues that within the context of curation, according to Duranti, the role of curators (or archivists) is to preserve evidence, not to determine the truth. The appraisal of data for intrinsic properties requires a high level of domain expertise regarding research methods, measurements, experiments, and potential novelties and impacts that are directly related to content. Thus, the capability of curators is another area of debate.

If the assessment of the intrinsic properties of data does not belong within the domain of traditional curation, it is important to think about how this assessment should be integrated into data curation either by re-defining the boundaries of curation or by collaborating with other professionals who can perform this role. Because good data curation practices encourage the involvement of and collaboration with different parties, the roles and responsibilities of data curation do not belong solely to data curators or data producers but to many other professionals as well. Thus, assessing the intrinsic properties of data may not be the responsibility of data

curators but may instead be tasked to other parties involved either before or during the curation life cycle.

Emerging discussions on data peer review have suggested one possible method for meeting users' expectations and sharing the responsibility of curating the intrinsic properties of data. Kratz and Strasser's (2015) survey, which examined studies from the sciences and social sciences, demonstrated that data peer review was more useful than any other factor in establishing the trustworthiness of data and in evaluating users' perceived value of the data. Given the effectiveness of data peer review, Mayernik et al.'s (2014) study results suggest the benefit of a potential collaboration between members of the scientific community, data producers, data curators, and users to conduct data peer reviews. Mayernik et al. (2014) sorted the properties of data into two categories: technical (e.g., metadata, documentation, file formats) and scientific (e.g., appropriate collection methods, validity, reliability). Mayernik et al. (2014) argued that it was essential to divide the data and to review these different aspects due to the different types of expertise required to review each aspect. For the intrinsic properties of data, experts in the scientific community should provide validation of the data, while data curators must concentrate on reviewing the technical aspects of the data. This division of responsibilities is an initial step for data to be properly managed by the original investigators (data producers) and to be understandable for potential reusers.

The present study also showed that potential (or future) data reusers had to contribute to the curation of intrinsic data properties. The participants of this study often noticed problems with the data during their own examination and use of the data. Some of these reusers would then conduct more thorough investigations on the data than the original investigators had performed, thus bringing a fresh set of eyes to the research situation. Reusers stimulated quality assurance

work on the data even after the data had passed the peer review. This role played by data reusers is integral not only for curating the intrinsic properties of data but also for all curation work conducted using the data.

4.5. Data reusers' awareness of the users' roles

This study also demonstrated the role played by users in data reuse and curation as well as the attributes needed to enhance or restore the trust these users had in the data (e.g., social acknowledgment, community reliance, and comprehensiveness). Several participants of this study remarked on the data contributed by the reusers during data management. For instance, PS15 preferred to use documentation of publically available data that had already been used by many researchers, as this documentation was more complete. This participant believed that the frequent use of this documentation helped develop more complete information about the data therein. Although the original investigators know all of the contextual information about the data, their views on the needs of reusers can be limited. The gap between what information is provided and what information is needed can be filled using the users' inputs. This is an important acknowledgement by the reusers of the users' roles and contributions to data management. At the same time, the fact that the data reusers' preferred data that were fulfilled by other users' needs suggests that data reusers would understand each other's needs based on their own individual experiences.

The final process of trust development also shows the data reusers' acknowledgment of other users' experiences. The data reusers searched for other reusers in various parties, including specific data reuser groups (if relevant groups exist), colleague groups, and the scholastic communities to which they belong. Other reusers can not only assist with the reusers' problem

178

solving but can also provide a feeling of reliance that can only be taken from a group experience, thus restoring the reusers' trust (which had previously been lowered). The reusers share experiences, and these shared experiences can be as influential on the data as the original investigators' authority.

This role played by the data reusers may be integral not only for curating the intrinsic properties of data but also for all curation work implemented on the data; however, the role of the data curators is also essential to the process of implementing users' input. Data curators can provide information regarding the users' needs and the expectations of the original investigators, thus contributing to achieving comprehensiveness and preparedness of the information. The curators also collaborate with the original investigators or help connect the original investigators with the users to find solutions to problems. Additionally, curators validate suspicious parts of the data reported by the reusers, document all of the processes, and maintain different versions of the data if newer versions are created. That being said, merely linking the data with any existing reuser group would also be an effective way to reinstate the trust that has already been built among the reusers.

5. Contribution

The first contribution of this research is to the study of trust. Researchers in various fields have been developing their own understandings and applications of trust specific to their respective domains, which may be one reason why there is such difficulty in defining and understanding trust across all the different disciplines. While it may not be realistic to expect one single definition of trust or one approach to understanding trust across all disciplines, the trust attributes from this study were used to show how users' trust in data can be conceptually mapped

179

with previous ideas regarding trust. New aspects of trust, which emerged in the context of data reuse, were also examined. The similarities between trust attributes across different research fields suggests the possibility of creating a high level of general agreement regarding trust understanding across different domains; however, the number of attributes specific to the data reuse context also suggests the need for new conditions for trust.

This research also offers several contributions to the fields of data reuse and data curation research. Although the literature on data reuse is continually growing, few attempts have been made to explain data reusers' behaviors from a theoretical perspective. While exploring the patterns and experiences of data reuse provides an understanding of data reuse practices, theoretical perspectives often lead to a deeper understanding, as this research analyzes the thoughts, perceptions, and beliefs behind the behaviors and actions of data reusers. Trust is a useful theoretical concept to explore in regard to data reusers' behaviors and perceptions; as explained in the first chapter, current data reuse practices make the concept of trust even more relevant to data reuse. By understanding data reusers' decision processes for selecting and reusing data for their research based upon their own trust judgments, this study's findings have confirmed the usefulness and significance of trust in data reuse. The theoretical understanding developed in this study has the potential to be applied to other contexts of data reuse through further studies.

By drawing upon a range of fields that concern the concept of trust—including sociology, social psychology, economics, information systems, and organizational behavior—this study also contributes to understanding the multiple facets of trust involved in the context of data reuse. A data reuser's trust judgment is not a one-time, simple process. Various types and levels of trust

interact to enable reusers to make trust judgments regarding data. Data reusers' trust in data is also not static; it can be changed at any time during the process of data reuse.

Data reusers' trust is involved in different levels of the data reuse process. Trust was not developed from a single entity or element in this study; rather, it was associated with a diverse range of social, institutional, individual, and object-level factors that consisted of multiple stages of trust development. The social factors involved in trust included the general practices and ethics of research, scholarly communities, and data reuser communities. Institutional factors encompassed research organizations and data repositories, while individual factors included the original investigators and other researchers. The object-level factors were the data and information about the data. All of these different types of factors worked together to build or diminish the data reusers' trust in the data. Ultimately, data reusers sought to trust in an object (data); however, the influence of these different attributes from all levels confirms that data are a social product in which multiple social parties are involved. The complexity of the data reusers' trust was shaped by the nature of the data as a social project. This complexity necessitated diverse approaches to the understanding of trust in a data reuse context.

This study also contributed to data curation research by providing implications for user-trusted data curation. The study presented the distinctive roles and impacts of different stakeholders or actors involved in data reuse, such as the original investigators (data producers), user communities, and perhaps even domain experts. Inviting these different parties to play roles in data curation could be important; the user-defined trust attributes cannot be associated with data curation without the contributions and active involvement of these different parties during the curation lifecycle. Although previous research in data curation has discussed the roles and responsibilities of these stakeholders (as seen in Table 3), the manner in which their roles

influence data reusers' experiences and judgment of the data has not yet been empirically well explained. In addition, data curation models, such as the DCC model (Figure 1), do not clarify the roles these stakeholders can play in the different activities of data curation throughout the curation lifecycle. Re-defining the expected roles of these stakeholders using the reusers' perspectives and clearly articulating the desired engagement of these stakeholders in the curation lifecycle would help curators meet reusers' expectations regarding the trustworthiness of data and their satisfaction with their experiences.

User-defined trust attributes provide some insights into the collection, preparation, and distribution of data and data information. Clearly, having more information was always better for the data reusers in this study, as they concerned themselves with "what" data or "what level (or detail) of" data information was collected. In addition to the "what" concerns, "how" or "in what manner" the data and data information were collected and prepared also mattered a great deal to the data reusers. They were additionally concerned with how the data and data information "look[ed]" to them when they received or acquired it. Together, these components comprised part of the data reusers' trust judgments. Although the integration of these users' needs and perspectives with curation practices still needs to be investigated further (e.g., what is the proper and comprehensive level of descriptions for data reusers? How can users' concerns be reflected in the ethics and transparency of using curation as a normal practice? How should data and data information be presented to reusers in terms of the best visuals and organizational methods?), learning more about reusers' thoughts and experiences is an important first step.

6. Limitations of the study

As with any research, this study has some limitations. More specifically, the limitations in this study were due to the research method, sampling, and study design. 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 due to the different forms and formats of data, the methods used to acquire the data, and the process of understanding and analyzing data. While this study informed the process of data reusers' development of trust using trust attributes in various contexts, the findings of this study may not be directly applicable to other types of data reusers, including qualitative, computational, and experimental data reusers.

The sampling strategy employed in this study also had its own limitation. Since the aim of this study was to provide details about the perceptions and understandings of a particular group, this study employed purposive sampling with a small sample size. Although purposive sampling is well justified in interpretative qualitative research, the major limitation of purposive sampling is that it produces a sample that may not represent a larger population, as probability sampling does. Thus, Glaser (1978) argued that in order to seek generalizability, the analysis of data collected from purposive sampling needs to provide a future direction for theoretical sampling.

Another limitation of this study was the process used to identify potential study participants. This method could be much improved in future studies on this topic if new citation tools appear. As explained in Chapter 3, study participants were manually selected from major scholarly databases due to a lack of fully established and implemented data citation standards. Although this method was very effective for identifying data reusers, the use of a manual search

had some limitations. First, the search results were limited to the journals and conference proceedings that were included in the databases used. Second, there was also a possibility that the authors of related publications represented in the citation databases may not use the terms "secondary data" or "secondary analysis," which were the terms used for the manual search. These two limitations may have inhibited access to other potential data reusers who could have been included in the study.

The final limitation of this study was that it mostly relied on the reported memories of participants' past experiences rather than on their current, ongoing data reuse experiences. This method was chosen because although investigating and observing data reusers' current processes of trust development through a field study in a natural setting would have provided more accurate information, this method also would have presented several obstacles and difficulties. Such obstacles included diversity of the geographic locations of the data reusers and difficulty in finding data reusers who were planning to start new projects using secondary data. Also, the starting point of data reuse is often blurred, as the process of data discovery is not always initiated with a new project. Human memories are imperfect and can be imprecise. Data reusers were identified by reviewing the most recent publications in reverse chronological order. This was done to find potential participants with recent data reuse experiences and to minimize the risk of memory recall errors. However, because the participants were encouraged to choose the experiences about which they liked to talk (not just the most recent experiences) and because most of the participants talked about multiple experiences from their pasts, the interview data still have the potential to be affected by memory imprecision.

7. Future studies

Several areas of future research could be pursued based on the findings of this research. First, comparative studies across different disciplines and data types are valuable. An investigation should be conducted to determine whether there are commonalities among user-defined trust attributes across various disciplines and data types because this research will guide the future development of data curation models by answering the following questions: What would be the essential trust attributes across different domains? Is there a need to build a data type-based curation model? If any differences between disciplines do appear, understanding the factors behind these differences should be pursued.

Future research should also be concerned with the development of trust measurement. Various trust attributes were defined in this study, but these attributes were only qualitatively discussed and could not be measured quantitatively. Even though this study presented changes in the level of users' trust, these changes were not explained quantitatively. Measuring trust will be essential in better detecting changes in trust and thus may determine what trust attributes account for those changes. Thus, measurement or scale development would be the next necessary step for quantifying data reusers' trust judgments and showing the changes in their trust levels under trustworthy data conditions.

APPENDICES

APPENDIX A. PARTICIPANTS DEMOGRAPHIC INFORMATION

	ID	Title	Field	Gender	Age	Years of experience
1	PH01	Assistant professor	Health	F	40s	15
2	PH02	Assistant professor	Health	F	40s	15
3	PH03	Research scientist	Health	F	40s	12
4	PH04	Post-doc	Health	M	20s	11
5	PH05	PhD student	Health	F	30s	6
6	PH06	Research scientist	Health	F	50s	20
7	PH07	Research scientist	Health	F	60s	35
8	PH08	Associate Professor	Health	F	60s	30
9	PH09	Research scientist	Health	F	30s	7
10	PH10-1	Professor	Health	F	60s	24
11	PH10-2	Research scientist	Health	M	30s	6
12	PH11	Associate professor	Health	F	40s	13
13	PH12	Assistant professor	Health	F	70s	45
14	PH13	Research scientist	Health	F	50s	25
15	PH14	Assistant professor	Health	F	30s	7
16	PH15	Assistant professor	Health	F	50s	25
17	PH16	Assistant professor	Health	F	30s	5
18	PH17	Assistant professor	Health	M	40s	12
19	PH18	Professor	Health	M	40s	16
20	SW01	Associate professor	Social work	F	40s	15
21	SW02	Professor	Social work	M	50s	33
22	SW03	Professor	Social work	F	60s	32
23	SW04	PhD student	Social work	F	30s	7
24	SW05	PhD student	Social work	F	30s	8
25	SW06	PhD student	Social work	F	30s	2
26	SW07	Assistant professor	Social work	M	30s	10
27	SW08	Research scientist	Social work	F	30s	12
28	SW09	Assistant professor	Social work	M	40s	10
29	SW10	Research scientist	Social work	F	30s	8
30	SW11	Research scientist	Social work	M	20s	5
31	SW12	Assistant professor	Social work	F	40s	11
32	SW13	Assistant professor	Social work	M	30s	13
33	SW14	PhD student	Social work	F	20s	4
34	SW15	Assistant professor	Social work	M	20s	6
35	SW16	PhD student	Social work	M	20s	5
36	SW17	Professor	Social work	M	50s	23

| 37 | SW18 | Professor | Social work | M | 50s | 26 |
| 38 | SW19 | Assistant professor | Social work | F | 30s | 9 |

APPENDIX B. SEMI-STRUCTURED INTERVIEW QUESTIONS

The researcher will use the following interview questions as starting points that help a participant to describe data reuser experiences.

Each interview will take approximately one hour. The researcher, however, will adjust the length of the interview based on a participant's comfort, attention span, and time constraint. The researcher may arrange follow-up interviews as initial interview data analysis moves along, depending on the participant's willingness to talk more about her experiences and topics and questions emerged during the first interview that stimulate further exploration.

Brief background of data reuse

- Gender
- Age group (20-29; 30-39; 40-49; 50-59; 60-69; 70 and older)
- Years of experience in the discipline
- Years of working with data (both working with own data as well as secondary data)
- Do you usually conduct *secondary analysis [localize the term in each discipline]* for your research, or do you also collect your own data?
    - Approximate frequency of collecting vs. *reusing [localize the term]* data
- On what occasions do you conduct *secondary analysis [localize the term]*?
    - How would you define data?
    - What do you perceive the role of data in research to be?

Data reuse experience (specific case)

Think about your past experience of *reusing [localize the term]* data for your research. Please pick one experience on which to focus your answers to the questions. You can also bring up other cases when you answer the questions.

- Can you briefly describe what the research project was about?
- What was the name of the data set you used for this research? What type of data were they?
- How did you find the data? (Searching process)
- How did you locate the data? (Access)
- How did you know the data was what you were looking for?
- What were the criteria for your decisions before actually start looking at and digging through the data?
    - (Probe) Was there any information of/about the data that you particularly liked to or need to check?
    - (Probe) Why did you think [x] information was important?
- What were the other steps before actually using the dataset?
- What were the next steps to understanding and using the data for your research?
- Any difficulties?
    - (Probe) How did you solve [any issue discussed]?

Trust and data

- How do you define the term trust? What do you mean by trust if you say "I trust xx"?
- How would you define (or characterize) trustworthy data?
- How did you decide if you could trust the dataset you used (regarding the dataset that you've been talking about)?
    - o (Probe) What kinds of information did you need in order to judge whether you could trust the data or not?
- Were there any changes on your trust judgment once you start playing with data?
    - o (Probe) In other words, during the process of cleaning, analyzing, and using data for your research, was there any change in your trust compared to a time when you just decided to use the data?)
- Have you experienced any difficulty judging how much you can trust certain data?
    - o (Probe) If so, what was the experience and why did that happen?
    - o (Probe) If so, what did you do?

Ending question

- What makes you lose your trust in data?
- Is there anything that you like to talk more about your data re-use *[localize the term]* experience and trust?
- Will you be able for a follow-up if I have a question or need a clarification?

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). doi: http://doi.org/10.1088/1742-6596/219/4/042018

Adelson, B. (1984). When novices surpass experts: The difficulty of a task may increase with expertise. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10*(3), 483-495. doi: 10.1037//0278-7393.10.3.483

Altheide, D., & Johnson, J. M. C. (1998). Criteria for assessing interpretive validity in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Collecting and Interpreting Qualitative Materials.* (pp. 283-312). Thousand Oaks, CA: Sage.

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.

Akmon, D. (2014). *The Role of Conceptions of Value in Data Practices: A Multi-Case Study of Three Small Teams of Ecological Scientists*. (Doctoral dissertation). Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/107162

Akmon, D., Zimmerman, A., Daniels, M., & Hedstrom, M. (2011). The application of archival concepts to a data-intensive environment: working with scientists to understand data management and preservation needs. *Archival Science, 11*(3-4), 329-348. doi:10.1007/s10502-011-9151-4

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

Anderson, E. & Weitz, B. (1989). Determinants of continuity in conventional industrial channel dyads. *Marketing Science, 8*(4), 310-323.

Angevaare, I. (2009). Taking care of digital collections and data: "Curation" and organizational choices for research libraries. *LIBER Quarterly*, *19*(1), 1–12.

Arena, R., Lazaric, N., & Lorenz, E. (2006). Trust, codification and epistemic communities: implementing an expert system in the French steel industry. In Reinhard and Akbar (Eds.), *Handbook of Trust Research*. (pp. 187-198). MA, Northamton: Edgar Elgar Publishing, Inc.

Baier, A. (1986). Trust and antitrust. *Ethics, 96*, 231-260.

Bailey K.D. (1987). *Methods* of *Social Research (*3rd ed.). New York: The Free Press.

Baker, J. (1987). Trust and Rationality. *Pacific Philosophical Quarterly*, 68, 1–13.

Baker, K. S., & Yarmey, L. (2008). Data stewardship: Environmental data curation and a web-of- repositories. *4th International Digital Curation Conference*, Edinburgh, Scotland, December, 2008.

Barber, B. (1983) *The Logic and Limits of Trust*. New Brunswick, NJ: Rutgers University Press.

Beagrie, N. (2008). Digital curation for science, digital libraries, and individuals. *International Journal of Digital Curation*, *1*(1), 3–16. doi:10.2218/ijdc.v1i1.2

Beagrie, N., Chruszcz, J., & Lavoie, B. (2008). *Keeping Research Data Safe: A Cost Model and Guidance for UK Universities*. London: JISC.

Bendiktsson, D. (1989). Hermeneutics: dimensions toward LIS thinking. *Library and Information Science Research, 11,* 201-234.

Berg, M., & Goorman, E. (1999). The contextual nature of medical information. *International Journal of Medical Informatics, 56*(51-6), 51-60. doi:10.1016/S1386-5056(99)00041-6

Bernard, H. R. (2000). *Social Research Methods: Qualitative and Quantitative Approaches*. Thousand Oaks, CA: Sage.

Blomqvist, K. (1997). The many faces of trust. *Scandinavian Journal of Management*, *13*(3), 271–286. doi:16/S0956-5221(97)84644-1

Birnholtz, J. P., & Bietz, M. (2003). Data at work: Supporting sharing in science and engineering. In *ACM Conference on Supporting Group Work* (pp. 339–348). Sanibel Island, FL.

Bishop, A. P. (1999). Document structure and digital libraries: How researchers mobilize information in journal articles. *Information Processing and Management, 35*(3), 255–279. doi:10.1016/S0306-4573(98)00061-2

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

Borgman, C. L. (2009). The digital future is now: A call to action for the humanities. *Digital Humanities Quarterly, 3*(4). Retrieved from http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html.

Borgman, C. L. (2010). Research data: Who will share what, with whom, when, and why? Presented at the *China-North America Library Conference*, Beijing. Retrieved from http://works.bepress.com/borgman/238

Borgman, C. L. (2011). The conundrum of sharing research data. *SSRN eLibrary*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1869155.

Borgman, C., Wallis, J. C., & Enyedy, N. (2006). Building digital libraries for scientific data: An exploratory study of data practices in habitat ecology. In the proceeding of *10th European Conference on Digital Libraries* (pp. 170–183). Springer. LINCS.

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007). Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*, *7*(1-2), 17–30. doi:10.1007/s00799-007-0022-9

Boslaugh, S. (2007). *Secondary Data Sources for Public Health: A Practical Guide*. Cambridge University Press. NY: New York.

Bowlby, J. (1982). *Attachment and loss, 1: Attachment* (2nd Ed.). New York: Basic Books (new printing, 1999, with a foreword by Allan N. Schore; originally published in 1969).

Brown, J. S., & Duguid, P. (2001). Knowledge and Organization: A Social-Practice Perspective. *Organization Science*, *12*(2), 198–213. Doi: 10.1287/orsc.12.2.198.10116

Budd, J. M. (1995). An epistemological foundation for library and information science. *Library Quarterly, 65*, 295-318.

Butler, J. K. (1991). Toward understanding and measuring conditions of trust: Evolution of a conditions of trust inventory. *Journal of Management*, *17*(3), 643–663. doi:10.1177/014920639101700307

Callaghan, S. (2015). Data without peer: Examples of data peer review in earth sciences. *D-Lib Magazine, 21*(1/2). doi: 10.1045/january2015-callaghan

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

Capurro, R. (2000). Hermeneutics and the phenomenon of information. In: Carl Mitcham, ed.: Metaphysics, Epistemology, and Technology. *Research in Philosophy and Technology, 19*, 79-85.

Carlson, S. (2006). Lost in a sea of science data. *The Chronicle of Higher Education, 23*. Retrieved from http://chronicle.com/article/Lost-in-a-Sea-of-Science-Data/9136

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

Castelfranchi, C., Falcone, R., & Pezzulo, G. (2003). Trust in information sources as a source for trust: a fuzzy approach. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, 89–96. doi:10.1145/860575.860590

Center for Research Libraries/Online Computer Library Center (CRL/OCLC). (2007). *Trustworthy Repositories Audit & Certification: Criteria and Checklist* (TRAC. Version 1.0).

Chen, L. (2005, September). A proof of concept: Provenance in a service oriented architecture. In *the proceedings of the UK e-Science All Hands Meeting*, Nottingham, UK.

Chesney, T. (2006). An empirical examination of Wikipedia's credibility. *First Monday*, *11*(11). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1413/1331

Chin, G., & Lansing, C. S. (2004). Capturing and supporting contexts for scientific data sharing via the biological sciences collaboratory. In *CSCW'04, ACM Conference on Computer Supported Cooperative Work* (pp. 409–418). Chicago, Illinois, USA: ACM.

Chopra, K., & Wallace, W. (2003). Trust in Electronic Environments. *Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03), 9*(9).

Clubb, J. M., Austin, E. W., Geda, C. L., & Traugott, M. W. (1985). Sharing research data in the social science. In S.E. Fienberg et al. (Eds.), *Commission on Behavioral and Social Sciences and Education* (pp. 39-88). Washington, D.C.: National Academy Press.

Coates, H. (2014). Ensuring research integrity: The role of data management in current crises. *College & Research Libraries News, 75*(11), 598-601.

Committee on Science, Engineering, and Public Policy (U.S.), Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, Committee on Policy and Global Affairs, Institute of Medicine, & National Academy of Sciences. (2009). *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. Washington, D.C.: National Academies Press.

Consultative Committee for Space Data Systems (CCSDS). (2011). *Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories* (No. CCSDS 652.1-M-1). The Consultative Committee for Space Data Systems.

Consultative Committee for Space Data Systems (CCSDS). (2002). *Reference Model for an Open Archival Information System (OAIS)* (No. CCSDS 650.0-B-1). The Consultative Committee for Space Data Systems.

Cornelius, I. (1996). *Meaning and Method in Information Science*. Norwood, NJ: Ablex.

Corti, L. (2000, December). Progress and problems of preserving and providing access to qualitative data for social research: The International Picture of an Emerging Culture. *Forum: Qualitative Social Research*, *1*(3). Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/1019/2197.

Corritore C, Kracher B, & Wiedenbeck S. (2003). On-line trust: Concepts, evolving themes, a

model. *International Journal of Human Computer Studies*, *58*, 737-58. doi: 10.1016/S1071-5819(03)00041-7

Costello, M. J. (2009). Motivating online publication of data. *BioScience*, *59*(5), 418–427. doi:10.1525/bio.2009.59.5.9

Cox, A. (2005). What are communities of practice? A comparative review of four seminal works. *Journal of Information Science*, *31*(6), 527–540. doi:10.1177/0165551505057016

Cragin, M. H., & Shankar, K. (2006). Scientific data collections and distributed collective practice. *Computer Supported Cooperative Work (CSCW)*, *15*(2-3), 185–204. doi:10.1007/s10606-006-9018-z

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023–4038. doi:10.1098/rsta.2010.0165

Data Governance and Quality: Data Reuse vs. Data Repurposing (2012). Retreived from http://dataqualitybook.com/?p=349#more-349

Data Seal of Approval (DSA). (n.d.). *Data Seal of Approval: Overview*. Retrieved from http://datasealofapproval.org/?q=about.

Data Seal of Approval (DSA). (2010). *Quality Guidelines for Digital Research Data*. Retrieved from http://datasealofapproval.org/?q=about.

Deutsch, M. (1962). Cooperation and trust: Some theoretical notes. In *Nebraska Symposium on Motivation, 1962* (pp. 275–320). Oxford, England: University of Nebraska Press.

Digital Curation Center (DCC). What is Digital Curation? (n.d.). Retrieved from http://www.dcc.ac.uk/digital-curation/what-digital-curation.

Digital Curation Center (DCC). DCC Curation Lifecycle Model. (n.d.). Retrieved from http://www.dcc.ac.uk/resources/curation-lifecycle-model.

Digital Preservation Coalition. (2009). *Preserving Geospatial Data*. York, UK: Digital Preservation Coalition. Retrieved from http://www.dpconline.org/component/docman/doc_download/363-preserving-geospatial-data-by-guy-mcgarva-steve-morris-and-gred-greg-janee-preserving-geospatial-data-by-guy-mcgarva-steve-morris-and-gred-greg-janee?q=geospatial+data.

Dingwall, G. (2004). Trusting archivists: The role of archival ethics codes in establishing public faith. *American archivist, 67*(1), 11-30. doi:abs/10.17723/aarc.67.1.mw0914r2p52xx2t4

Donaldson, D. R. & Conway, P. (2015). User conceptions of trustworthiness for digital archival documents. *Journal of the Association for Information Science and Technology*. Article first published online: 30 JAN 2015. doi: 10.1002/asi.23330

Donaldson, D. R. & Fear, K. (2011). Provenance, End-User Trust, and Reuse: An Empirical Investigation. In the *Proceedings of the 3rd USENIX Workshop on the Theory and Practice of Provenance* (TaPP'2011) Heraklion, Crete, Greece.

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

Duerr, R., Parsons, M. A., Marquis, M., Dichtl., R., & Mullins, T. (2004). Challenges in long-term data stewardship. Paper presented at the *12th NASA Goddard/21st IEEE Conference on Mass Storage Systems and Technologies* (pp. 101-121). University of Maryland, MD.

Elliott, R., Fischer, C. T., & Rennie, D. L. (1999). Evolving guidelines for publication of qualitative research studies in psychology and related fields. *British Journal of Clinical Psychology*, *38*(3), 215–229. doi:10.1348/014466599162782

Erikson, E. H. (1964). *Childhood and Society*. New York, NY: Norton.

Erwin, T., Sweetkind-Singer, J., & Larsgaard, M. (2009). The national geospatial digital archives—collection development: Lessons learned. *Library Trends*, *57*(3), 490–515.

ESRC Research Data Policy. (2010). Economic and Social Research Council. Retrieved from http://www.esrc.ac.uk/about-esrc/information/data-policy.aspx.

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., & Zimmerman, A. (2011). Beyond the data deluge: A research agenda for large-scale data sharing and reuse. *International Journal of Digital Curation*, *6*(1). doi:10.2218/ijdc.v6i1.172

Faulkner, P. (2010). Our norms of trust. In A. Haddock, A. Millar & D. Pritchard (Eds.), *Social Epistemology*. (pp. 129-147). Oxford, U.K.: Oxford University Press.

Fear, K. (2013). Measuring and Anticipating the Impact of Data Reuse. (Doctoral dissertation). Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/102481

Flanagin, A., & Metzger, M. (2000). Perceptions of Internet information credibility. *Journalism & Mass Communication Quarterly, 77*(3), 515-540. doi: 10.1177/107769900007700304

Flick, U. (2009). *An Introduction to Qualitative Research*. London, UK: Sage.

Fritch, J., & Cromwell, R. (2001). Evaluating Internet resources: Identity, affiliation, and cognitive authority in a networked world. *Journal of American Society of Information Science Technology, 52*(6), 498–507.

Fienberg, S.E. Martin, M .E. & Straf, M .L., National Research Council(U.S.) Committee on National Statistics., & National Research Council (U.S.). Commission on Behavioral and Social Sciences and Education (1985). *Sharing Research Data*. Washington, D.C.: National Academy Press.

Fulmer, A. C. & Gelfand, M. J. (2013) How do I trust thee? Dynamic trust patterns and their individual and social contextual determinants. In Models for Intercultural Collaboration and Negotiation, K. Sycara et al. (Eds). Advances in Group Decision and Negotiation 6. doi: 10.1007/978-94-007-5574-1_5

Fukuyama, F. (1996). *Trust: The Social Virtues and The Creation of Prosperity*. New York, NY: Simon and Schuster.

Galliers, R. D., & Newell, S. (2003). Back to the future: from knowledge management to the management of information and data. *Information Systems and e-Business Management, 1*(1), 5–13. doi:10.1007/BF02683507

Gambetta, D. (1988). Can we trust trust? In D. Gambetta (Ed.), *Trust: Making and Breaking Cooperative Relations* (pp. 213-238). Oxford: Basil Blackwell.

Glaser B. G. (1978). Theoretical Sensitivity. Mill Valley, CA: The Sociology Press.

Glaeser, P. S. (1990). *Scientific and Technical Data in a New Era*. New York, NY: Hemisphere Publishing Corporation.

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

Giddens, A. (1990). *The Consequences of Modernity*. Stanford, CA: Stanford University Press.

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.

Gitelman, L., Ed. (2013). *"Raw Data" is an Oxymoron*. Cambridge: MA: MIT Press.

Gold, A. K. (2013). Libraries, Process, and Data. In *Proceedings of the Association for Information Science and Technology, 50*(1), 1-9.

Good, D. (1988). Individuals, interpersonal relations and trust. In D. Gambetta (Ed.), *Trust: Making and Breaking Cooperative Relations* (pp. 31–48). Oxford, U.K.: 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. doi:10.1117/12.461524

Green, A. G., & Gutmann, M. P. (2007). Building partnerships among social science researchers, institution-based repositories and domain specific data archives. *OCLC Systems & Services*, *23*(1), 35–53. doi:10.1108/10650750710720757

Gregory, A. (July 2011). *The Data Documentation Initiative (DDI): An Introduction for National Statistical Institutes*. Open Data Foundation. Retrieved from http://odaf.org/papers/DDI_Intro_forNSIs.pdf.

Groth, P. (2005, September). PReServ: Provenance recording for services. Paper presented at the *UK e-Science All Hands Meeting*, Nottingham, UK.

Guest, G. & Namey, E. E. (2014). *Public Health Research Methods*. Thousand Oaks, CA: SAGE publications.

Gutmann, M., Schurer, K., Donakowski, D., & Beedham, H. (2004). The selection, appraisal and retention of social science data. *CODATA Data Science Journal, 3*, 209-221. doi: http://doi.org/10.2481/dsj.3.209

Haas, P. M. (1992). Introduction: epistemic communities and international policy coordination. *International Organization*, *46*(01), 1–35. doi: http://dx.doi.org/10.1017/S0020818300001442

Halinen, A. (1994) Exchange Relationships in Professional Services. A Study of Relationship Development in the Advertising Sector. (Doctoral dissertation). Turku School of Economics and Business Administration.

Hansson, J. (2005). Hermeneutics as a bridge between the modern and the postmodern in library and information scienc. *Journal of Documentation, 61*(1). 102-113. doi: http://dx.doi.org/10.1108/00220410510578032

Hedstrom, M., & Niu, J. (2008). Research forum presentation: Incentives to create "Archive-Ready" data: Implications for archives and records management. In the *Proceedings of the American Archivist(SAA) research forum*.

Heidorn, P. B. (2011). The emerging role of libraries in data curation and E-Science. *Journal of Library Administration, 51*(7), 662–72. doi:10.1080/ 01930826.2011.601269.

Hertzberg, L. (1988). On the attitude of trust. *Inquiry, 31*(3), 307-322. doi: 10.1080/00201748808602157

Hertzum, M., Andersen, H. H., Andersen, V., & Hansen, C. B. (2002). Trust in information sources: seeking information from people, documents, and virtual agents. *Interacting with Computers*, *14*(5), 575–599. doi:10.1016/S0953-5438(02)00023-1

Hey, T., Tansely, S. & Tolle, K. (Eds.). (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/4thparadigm_science.pdf.

Hey, T., Trefethen, A. (2003). The data deluge: An e-science perspective. In F. Berman, G.C. Fox, & T. Hey, (Eds.), *Grid computing: Making the global infrastructure a reality*. New York, NY: Wiley.

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation*, *3*(1), 134–140. doi:10.2218/ijdc.v3i1.48

Hilgartner, S. & Brandt-Rauf, S. I. (1994). Data access, ownership, and control. *Knowledge: Creation, Diffusion, Utilization* 15(4), 355-372. doi: 10.1177/107554709401500401

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). PA, Hershey: Idea Group Publishing.

Ho, L. A., Kuo, K. T., Lin, C., & Lin, B. (2010). The mediate effect of trust on organizational online knowledge sharing: an empirical study. *International Journal of Information Technology & Decision Making, 9*(4), 625-44. doi: 10.1142/S0219622010003981

Hollander, E. P. (1958). Conformity, status, and idiosyncrasy credit. *Psychological Review, 65*(2). 117-127. doi: http://dx.doi.org/10.1037/h0042501

Hovland, C. I. , Harvey, O. J., & Sherif, M. (1957). Assimilation and contrast effects in reactions to communication and attitude change. *Journal of Abnormal and Social Psychology, 55*(2), 244-252. doi: http://dx.doi.org/10.1037/h0048480

Humphrey, C. K., Estabrooks, C. A., Norris, J. R., Smith, J. E., & Hesketh, K. L. (2000). Archivist on board: Contributions to the research team. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *1*(3). Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/1022/220.

Hutchinson, S., & Skodol-Wilson, H. (1992). Validity threats in scheduled semi-structured research interviews. *Nursing Research* 41(2), 117-119.

Inter-university Consortium for Political and Social Research (ICPSR). (2002). *Guide to Social Science Data Preparation and Archiving*. Inter-university Consortium for Political and Social Research. Retrieved from http://www.icpsr.umich.edu/access/dataprep.pdf.

Inter-university Consortium for Political and Social Research (ICPSR). (2009). *Principles and Good Practice for Preserving Data* (IHSN Working Paper No. no 003). ICPSR. Retrieved from http://www.ihsn.org/home/sites/default/files/resources/IHSN-WP003.pdf

Jones, G. R., & George, J. M. (1998). The experience and evolution of trust: Implications for cooperation and teamwork. *The Academy of Management Review*, *23*(3), 531–546. doi:10.2307/259293

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

Joint Information Systems Committee (JISC). (2003). *An invitation for expressions of interest to establish a new Digital Curation Centre for research into and support of the curation and preservation of digital data and publications*. Retrieved at http://www.jisc.ac.uk/fundingopportunities/funding_calls/2003/09/funding_digcentre.aspx

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

Key Perspectives Ltd. (2010). *Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long term Viability-A comparative review based on sixteen case studies*. Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/sites/default/files/SCARP%20SYNTHESIS_FINAL.pdf

King, G. (1995). Replication, Replication. *PS: Political Science and Politics, 28*(3), 444–452. doi:10.2307/420301

Kini, A. (1998). Trust in Electronic Commerce: Definition and Theoretical Considerations. In the *Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences,* 4, 51-62.

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

Kvale, S. (1996). *Interviews: An introduction to Qualitative Research Interviewing*. Thousand Oaks, CA: Sage

Kunze, J., Cruse, P., Hu, R., Abrams, S., Hastings, K., Mitchell, C., & Schiff, L. R. (2011). *eScholarship: Practices, Trends, and Recommendations in Technical Appendix Usage for*

Selected Data-Intensive Disciplines. Retrieved from
http://escholarship.org/uc/item/9jw4964t.

Lagerspetz, O. (1992). Legitimacy and trust. *Philosophical Investigations, 15*, 1-21.

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

Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011). Citation and peer
review of data: Moving towards formal data publication. *International Journal of Digital
Curation*, *6*(2), 4–37. doi:10.2218/ijdc.v6i2.205

Lazaric, N. (2003). Trust building inside the 'epistemic community': an investigation with an
empirical case study. In Nooteboom & Six (Eds.). *The Trust Process in Organizations:
Empirical Studies of the Determinants and the Process of Trust Development*. (pp. 147-
167). Northamton, MA: Edgar Elgar Publishing, Inc.

Lemieux, V. (2014). *Why We're Failing to Get the Most Out of Open Data*. Retrieved from
http://www.ideaslaboratory.com/post/99984665378/why-were-failing-to-get-the-most-
out-of-open-data.

Leininger, M. (1985) *Qualitative Research Methods in Nursing*. Orlando, FL: Grune & Stratton.

Leininger, M. (1994). Evaluation criteria and critique of qualitative research studies. In J. M.
Morse (Ed.), *Critical Issues in Qualitative Research Methods*. Newbury Park, CA: Sage.

Lewicki, R., & Bunker, B. (1996). Developing and maintaining trust in work relationships. In R.
M. Kramer & T. R. Tyler (Eds.), *Trust in Organizations: Frontiers of Theory and
Research* (pp. 114–139). Thousand Oaks, CA: Sage Publications.

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

Lincoln, Y. & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

Lord, P., & Macdonald, A. (2003). *e-Science Curation Report- Data curation for e-Science in
the UK: an audit to establish requirements for future curation and provision*. A report
prepared for the JISC Committee for the Support of Research (JCSR). Twickenham, UK:
The Digital Archiving Consultancy Limited.

Lord, P., Macdonald, A., Lyon, L., & Giarretta, D. (2004). From data deluge to data curation. In
UK e-science All Hands meeting 2004 (pp. 371–375). Retrieved from
http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/150.pdf.

Louis, K. S., Jones, L. M. & Campbell, E. G. (2002). Macroscope: Sharing in Science. *American
Scientist 90* (4), 304-307.

Lorenz, E. H. (1988). Neither Friends nor Strangers: Informal Networks of Subcontracting in

French Industry. In D. Gambetta (Ed.), *Trust: Making and Breaking Cooperative Relations* (pp. 194-210). Oxford, U.K.: Basil Blackwell.

Lorenz-Meyer, D. (2009). Possibilities of Enacting and Researching Epistemic Communities. *Sociological Research Online*, *15*(2), 13.

Lucassen, T., & Schraagen, J. M. (2011). Factual accuracy and trust in information: The role of expertise. *Journal of the American Society for Information Science and Technology, 62*(7), 1232-1242. doi:10.1002/asi.21545

Luhmann, N. (1979). *Trust and power*. Chichester, U.K.: Wiley.

Lyon, L. (2007). *Dealing with Data: Roles, Rights, Responsibilities, and Relationships*. UKOLN, University of Bath. Retrieved from http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2005/dealingwithdata#downloads.

Magrath, A. J., & Hardy, K. G. (1989). A strategic paradigm for predicting manufacturer-reseller conflict. *European Journal of Marketing*, *23*(2), 94–108. doi:10.1108/EUM0000000000549

Markus, M. L. (2001). Toward a theory of knowledge reuse: type of knowledge reuse situations and factors in reuse success. *Journal of Management Information Systems. 18*(1), 57-93.

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

Martinez-Uribe, L., & Macdonald, S. (2009). User engagement in research data curation. In *Proceedings of the 13th European conference on Research and advanced technology for digital libraries* (pp. 309–314). Berlin, Heidelberg: Springer-Verlag.

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

Mayernik, M. S., Wallis, J. C., Pepe, A., & Borgman, C. L. (2008). Whose data do you trust? Integrity issues in the preservation of scientific data. Paper presented at *iConference* 2008.

Mayernik, M. S., Callaghan, S., Leigh, R., Tedds, J., & Worley, S. (2014). Peer Review of Datasets: When, Why, and How. *Bulliten of American Meteorological Society*, 191-201.

McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *The Academy of Management Journal*, *38*(1), 24–59. doi:10.2307/256727

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

McDonough, J. (2012). Start making sense: quality, context & mining. In *Curating for Quality: Ensuring Data Quality to Enable New Science* (Workshop Report) (pp. 20-31). Retrieved from http://datacuration.web.unc.edu/.

McKnight, D. H., & Chervany, N. L. (1996). *The meanings of trust* (Technical Report No. 96-04). Minneapolis, MN: University of Minnesota.

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *The Academy of Management Review*, *23*(3), 473–490. doi:10.2307/259290

McKnight, D. H., & Kacmar, C. (2006). Factors of Information credibility for an Internet advice site. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences, 2006. HICSS '06, 6*. doi: 10.1109/HICSS.2006.181

Meyer, M., & Molyneux-Hodgson, S. (2010). Introduction: The Dynamics of Epistemic Communities. *Sociological Research Online*, *15*(2), 14.

Meyerson. D., Weick, K. E,. & Kramer, H. M. (1996). Swift trust and temporary groups. In R. M. Kramer & T. R. Tyler (Eds.), *Trust In Organizations: Frontiers of Teory and Research* (pp. 166-195). Thousand Oaks, CA: Sage.

Miles, R. E. & Snow, C. C. (1992). Causes of Failure in Network Organizations. *California Management Review*, summer 1992, 53-72.

Mishra, A. K. (1996). Organizational response to crisis: The centrality of trust. In R. M. Kramer & T. R. Tyler (Eds.), *Trust in Organizations: Frontiers of Theory and Research* (pp. 261-287). Thousand Oaks, CA: Sage

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

Morgan, R. M., & Hunt, S. D. (1994). The commitment-trust theory of relationship marketing. *Journal of Marketing*, *58*(3), 20–38. doi:10.2307/1252308

Morrow, S. L. (2005). Quality and trustworthiness in qualitative research in counseling psychology. *Journal of Counseling Psychology*, *52*(2), 250–260. doi:10.1037/0022-0167.52.2.250

Morse J.M., Barrett M., Mayan M., Olson K. & Spiers J. (2002) Verification strategies for establishing reliability and validity in qualitative research. *International Journal of Qualitative Methods, 1*(2) 1–19.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905–1922. doi:10.1080/00140139408964957

National Academy of Science. (2009). *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age.* Washington, DC: NAS. Retrieved from http://www.nap.edu/catalog.php?record_id=12615.

National Institutes of Health (NIH). (2003). *Data sharing policy: Final NIH Statement on Sharing Research Data*. Retrieved from http://grants2.nih.gov/grants/policy/data_sharing/.

National Research Council. (1999). *A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases*. Washington, DC: National Academy Press.

National Research Council. (2007) *Environmental Data Management at NOAA: Archiving, Stewardship, and Access.* Washington, DC: National Academies Press.

National Science Board. (2005). *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century* (No. NSB-05-40). National Science Foundation. Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/.

National Science Foundation (NSF). (2010). *Dissemination and Sharing of Research Results*. Retrieved from http://www.nsf.gov/bfa/dias/policy/dmp.jsp.

Nestor Working Group on Trusted repositories Certification. (2006). *Catalogue of criteria for trusted digital repositories: version 1*. Retrieved from http://files.d-nb.de/ nestor/materialien/nestor_mat_08-eng.pdf

Niu, J. (2009). Overcoming Inadequate Documentation. In *Proceedings of the Annual Meeting of the American Society for Information Science & Technology (ASIS&T), 46*(1). doi: 10.1002/meet.2009.145046024

Niu, J., & Hedstrom, M. (2008). Documentation Evaluation Model for Social Science Data. *Proceedings of the Annual Meeting of the American Society for Information Science & Technology (ASIS&T), 14*(1). doi: 10.1002/meet.2008.1450450223

Orlikowski, W.J. & Baroudi, J.J. (1991). Studying Information Technology in Organizations: Research Approaches and Assumptions. *Information Systems Research* (2), 1-28. doi: http://dx.doi.org/10.1287/isre.2.1.1

Packer, M. J. & Addison, R. B. (1989). Evaluating an interpretive account. In M. J. Packer & R. B. Addison, *Entering the circle: Hermeneutic investigation in psychology*, (pp. 275-292). Albany, NY: SUNY Press.

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

Patel, M., & Ball, A. (2008). Challenges and issues relating to the use of representation information for the digital curation of crystallography and engineering data. *International Journal of Digital Curation*, *3*(1), 76–88. doi:10.2218/ijdc.v3i1.43

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Polkinghorne, D. E. (1989). Phenomenological research methods. In R.S. Valle & S. Halling (Eds.) *Existential-phenemenological perspectives in psychology* (pp. 41-60). New York, NY: Plenum.

Popay J., Rogers A. & Williams G. (1998) Rationale and standards for the systematic review of qualitative literature in health services research. *Qualitative Health Research 8* (3), 341–351. doi: 10.1177/104973239800800305

Poutanen, S. (2001). How could contemporary social theory contribute to socialize epistemology? *Social Epistemology, 15*, 27–41. doi: 10.1080/02691720110049224

Prieto, A. G. (2009). From conceptual to perceptual reality: trust in digital repositories. *Library Review*, *58*(8), 593-606. doi:10.1108/00242530910987082

Pirson, M., & Malhotra, D. (2011). Foundations of Organizational Trust: What Matters to Different Stakeholders? *Organization Science, 22*(4), 1087-1104. doi: 10.1287/orsc.1100.0581

Radford, G. P. (1992). Positivism, Foucault, and the Fantasia of the Library: Conceptions of Knowledge and the Modern Library Experience. *Library Quarterly, 62*(October 1992), 408-424.

Research Libraries Group/Online Computing Library Center (RLG/OCLC) Working Group on Digital Archive Attributes. (2002). *Trusted Digital Repositories: Attributes and Responsibilities*. Retrieved from http://www.rlg.org/longterm/repositories.pdf

Research Information Network (RIN). (2008). *Stewardship of digital research data: A framework of principles and guidelines*. Retrieved from http://www.rin.ac.uk/our-work/data-management-and-curation/stewardship-digital-research-data-principles-and-guidelines.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, *49*(1), 95–112. doi:10.1037/0022-3514.49.1.95

Renzl, B. (2008). Trust in management and knowledge sharing: the mediating effects of fear and knowledge documentation. *Omega, 36*(2), 206-220.

doi:10.1016/j.omega.2006.06.005

Rieh, S. Y. & Belkin, N. J. (1998). Understanding Judgment of Information Quality and Cognitive Authority in the WWW. In the *Proceedings of the 61st Annual Meeting of the American Society for Information Science, 35*, 279-289.

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

Rolfe, G. (2006). Validity, trustworthiness and rigour: quality and the idea of qualitative research. *Journal of Advanced Nursing*, *53*(3), 304–310. doi:10.1111/j.1365-2648.2006.03727.x

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust1. *Journal of Personality*, *35*(4), 651–665. doi:10.1111/j.1467-6494.1967.tb01454.x

Rotter, J. B. (1971). Generalized expectancies for interpersonal trust. *American Psychologist, 26*(5), 443-452. doi: 10.1037/h0031464

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

Rusbridge, C., Burnhill, P., Ross, S., Buneman, P., Giaretta, D., Lyon, L., & Atkinson, M. (2005). The digital curation centre: A vision for digital curation. In *In proceeding of: Local to Global Data Interoperability - Challenges and Technologies* (pp. 31– 41). Forte Village Resort, Sardinia, Italy. doi:10.1109/LGDI.2005.1612461

Sako, M. (1992). *Prices, Quality and Trust: Inter-Firm Relations in Britain and Japan*. Cambridge, U.K.: Cambridge University Press.

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

Sandusky, R. J., & Tenopir, C. (2007). Finding and using journal article components: Impacts of disaggregation on teaching and research practice. *Journal of the American Society of Information Science and Technology, 59*(6), 970–982. doi: 10.1002/asi.20804

Schwartz-Shea, P. (2012). *Interpretative Research Design: Concepts and Processes*. New York, NY: Routledge.

Shankar, K. (2007). Order from chaos: The poetics and pragmatics of scientific record keeping. *Journal of the American Society for Information Science and Technology 58,* (10), 1457-1466. doi: 10.1002/asi.v58:10

Sharing Data from Large-scale Biological Research Project: A System of Triparties Responsibility. (2003). Meeting organized by the Wellcome Trust, Fort Lauderdale, Florida, Wellcome Trust. Retrieved from http://www.wellcome.ac.uk/About-us/Publications/Reports/Biomedical-science/WTD003208.htm.

Sheppard, B. H., & Sherman, D. M. (1998). The grammars of trust: A model and general implications. *Academy of Management Review*, *23*(3), 422–437. doi:10.5465/AMR.1998.926619

Sherif, M. & Hovland, C. I. (1961). *Social judgment: Assimilation and contrast effects in communication and attitude change.* Oxford, U.K.: Yale University Press.

Shorish, Y. (2012). Data curation is for everyone! The case for master's and baccalaureate institutional engagement with data curation. *Journal of Web Librarianship*, *6*(4), 263–273. doi:10.1080/19322909.2012.729394

Shreeves, S. L., & Cragin, M. H. (2008). Introduction: Institutional repositories: Current state and future. *Library Trends, 57*(2), 89–97. doi:10.1353/lib.0.0037

Sieber, J. (1991). Social scientists' concerns about sharing data. In John E. Sieber (Ed.). *Sharing Social Science Data*. (pp. 141-150). Thousand Oaks, CA: Sage publications.

Sitkin, S. B., & Roth, N. L. (1993). Explaining the limited effectiveness of legalistic "Remedies" for trust/ distrust. *Organization Science*, *4*(3), 367–392. doi:10.2307/2634950

Smith, J. A. (1996). Beyond the divide between cognition and discourse: Using interpretative phenomenological analysis in health psychology. *Psychology & Health*, *11*(2), 261–271. doi:10.1080/08870449608400256

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). 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

Smith, J. A., & Osborn, M. (2008). Interpretative Phenomenological Analysis. In *Qualitative psychology : a practical guide to research methods* (pp. 53–80). Los Angeles, CA: Sage.

Smith, V. S. (2009). Data publication: towards a database of everything. *BMC Research Notes*, *2*(1), 113. doi:10.1186/1756-0500-2-113

Special issue on qualitative data. (2010). *IASSIST Quarterly, 34*(3&4). Retrieved from http://www.iassistdata.org/iq/issue/34/3

Special issue, Qualitative Inquiry: Research, Archiving, and Reuse. (2005). *Forum: Qualitative*

Social Research, 6(2). Retrieved from http://www.qualitative-research.net/index.php/fqs/issue/view/12.

Speck, J. (2010). Protecting public trust: An archival wake-up call. *Journal of Archival Organization, 8*(1), 31-53. doi: 10.1080/15332748.2010.483389

Steinhart, G., Saylor, J., Albert, P., Alpi, K., Baxter, P., Brown, E., Chang, K., Corson-Rikert, J., Hirtle, P., Jenkins, K., Lowe, B., McCue, J., Ruddy, D., Silterra, R., Solla, L., Stewart-Marshall, Z., & Westbrooks, E. L. (2008). *Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library*. Retrieved from http://ecommons.library.cornell.edu/handle/1813/10903.

Stewart, L. (1996). User acceptance of electronic journals: Interviews with chemists at Cornell University. *College & Research Libraries, 57*(4), 339–349.

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

Sztompka, P. (1999). *Trust: A Sociological Theory*. Cambridge, U.K.: Cambridge University Press.

Tomlinson, E. C. & Mayer, R. C. (2005). The role of casual attribution dimensions in trust repair. *The Academy of Management Review, 34*(1). 85-104. doi: 10.5465/AMR.2009.35713291

Trust. (n.d.). Retrieved July 7, 2015, from http://www.merriam-webster.com/dictionary/trust

Tschannen-Moran, M. (2001). Collaboration and the need for trust. *Journal of Educational Administration*, 39, 308–331.

Tseng, S., & Fogg, B. (1999). Credibility and computing technology. *Communications of ACM*, *42*(5), 39–44. doi:10.1145/301353.301402

UK Data Archive. (2011). *Data Management Recommendations: For Research Centres and Programmes*. Retrieved from http://www.data-archive.ac.uk/media/257765/ukda_datamanagementrecommendations_centresprogrammes.pdf

Van den Berg, H. (2005). Reanalyzing Qualitative Interviews from Different Angles: The Risk of Decontextualization and Other Problems of Sharing Qualitative Data. *Forum: Qualitative Social Research*, *6*(1). Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/499/1074.

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

Van House, N.A. (2003). Digital libraries and collaborative knowledge construction. In A. P. Bishop, B. Buttenfield, & N. A. Van House, (Eds.), *Digital library use: Social practice in design and evaluation*. Cambridge, MA: MIT Press.

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

van Manen, M. (1997). *Researching lived experience: Human science for an action sensitive pedagogy* (2nd Ed.)*.* London, Canada: The Althouse Press.

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. *European Conference on Research and Advanced Technology for Digital Libraries*, Budapest, Hungary, 2007

Wallis, J. C., Borgman, C. L., Mayernik, M. S., & Pepe, A. (2008). Moving Archival Practices Upstream: An Exploration of the Life Cycle of Ecological Sensing Data in Collaborative Field Research. International Journal of Digital Curation, 3(1), 114-126.

Walsham, G. (1993). *Interpreting Information Systems in Organizations*. Chichester, U.K.: Wiley.

Walsham, G. (1995). Interpretive case studies in IS research: nature and method. *European Journal of Information Systems*, 4, 74-81. doi:10.1057/ejis.1995.9

Walsham, G. (1995). The Emergence of Interpretivism in IS Research. *Information Systems Research, 6*(4), 376-394. doi: http://dx.doi.org/10.1287/isre.6.4.376

Weber, J. M., Malhotra, D., & Murnighan, J. K. (2004). Normal acts of irrational trust: Motivated attributions and the trust development process. *Research in Organizational Behavior*, *26*, 75–101. doi:10.1016/S0191-3085(04)26003-8

Wellcome Trust. (2001). *Wellcome Trust Policy on Access to Bioinformatics Resources by Trust-Funded Researchers*. Retrieved from http://www.wellcome.ac.uk/doc%5Fwtd002759.html.

Wellcome Trust statement on genome data release. (1997). Retrieved from http://www.wellcome.ac.uk/doc%5Fwtd002751.html.

Welman, J. C., & Kruger, S. J. (1999). *Research Methodology for the Business and Administrative Science*s. Johannesburg, South Africa: International Thompson.

Wenger, E. (1998). *Communities of Practice: Learning, Meaning, and Identity*. Cambridge, U.K.: Cambridge University Press.

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, U.K.: Open University Press.

Willis, J. W. (2007). *Foundations of Qualitative Research: Interpretative and Critical Approaches*. Thousand Oaks, CA: Sage Publications.

Worchel, P. (1979). Trust and distrust. In W.G. Austin & S. Worchel (Eds.), *The Social Psychology of Intergroup Relations*. Belmont, CA: Wadsworth.

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

Zand, D. (1972). Trust and managerial problem solving. *Administrative Science Quarterly*. *17*, 229-239.

Zimmerman, A. S. (2003). *Data sharing and secondary use of scientific data: Experiences of ecologists* (Doctoral dissertation). Retrieved from http://search.proquest.com.libproxy.lib.unc.edu/docview/287907131/abstract?accountid=14244.

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

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 Organizational Behavior, 8*, 55-111.