Data curation for qualitative data reuse and big social research: Connecting communities of practice

D i s s e r t a t i o n

zur Erlangung des akademischen Grades

Doctor philosophiae
(Dr. phil.)

eingereicht

an der Philosophischen Fakultät
der Humboldt-Universität zu Berlin

von Sara Mannheimer

Der Präsident der Humboldt-Universität zu Berlin
Prof. Peter A. Frensch, PhD

Der Dekan der Philosophischen Fakultät
Prof. Dr. Thomas Sandkühler

Gutachter/innen

Erstgutachterin: Prof. Vivien Petras, PhD
Zweitgutachter: Prof. Michael Zimmer, PhD

Datum der Disputation: 22. Juli 2022

Abstract

Trends toward open science practices, along with advances in technology, have promoted increased data archiving in recent years, thus bringing new attention to the reuse of archived qualitative data. Qualitative data reuse can increase efficiency and reduce the burden on research subjects, since new studies can be conducted without collecting new data. Qualitative data reuse also supports larger-scale, longitudinal research by combining datasets to analyze more participants. At the same time, qualitative research data can increasingly be collected from online sources. Social scientists can access and analyze personal narratives and social interactions through social media such as blogs, vlogs, online forums, and posts and interactions from social networking sites like Facebook and Twitter. These big social data have been celebrated as an unprecedented source of data analytics, able to produce insights about human behavior on a massive scale. This study addresses the following research questions. **RQ1:** How is big social data curation similar to and different from qualitative data curation? **RQ1a:** How are epistemological, ethical, and legal issues different or similar for qualitative data reuse and big social research? **RQ1b:** How can data curation practices such as metadata and archiving support and resolve some of these epistemological and ethical issues? **RQ2:** What are the implications of these similarities and differences for big social data curation and qualitative data curation, and what can we learn from combining these two conversations? My research employs a social constructivist paradigm and engages with the theories of communities of practice and epistemic cultures. I answered my research questions through an in-depth review of the literature and semi-structured interviews. In the literature review, I identified six key issues in common between qualitative data reuse and big social research. The semi-structured interviews were conducted with three distinct communities of practice: qualitative researchers, big social researchers, and data curators. I used critical incident technique to structure the interview guide and I followed grounded theory methodology for conducting a qualitative content analysis of interview transcripts. This dissertation research produced the following insights. First, this research identified six key issues for qualitative data reuse and big social research: context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. Second, this research showed that each community of practice—qualitative researchers, big social

researchers, and data curators—viewed each of the six issues through a different lens, thus prioritizing different dimensions of each issue. This variation in perspective shows that connecting these three communities of practice can support a broader understanding of the key issues, and therefore lead to more epistemologically sound, ethical, and legal research practices. Third, this dissertation finds that data curators are well-positioned to provide guidance for epistemologically sound, ethical, and legal qualitative data reuse and big social data. Curators have the skills and perspectives to translate between communities of practice, and they have the competencies to take care of both types of data.

Zusammenfassung

Trends in Richtung Open-Science-Praktiken haben zusammen mit technologischen Fortschritten in den letzten Jahren eine verstärkte Datenarchivierung gefördert und damit der Nachnutzung archivierter qualitativer Daten neue Aufmerksamkeit geschenkt. Nachnutzung von qualitativen Daten (qualitative data re-use) kann die Effizienz steigern und die untersuchten Populationen entlasten, da neue Studien durchgeführt werden können, ohne neue Daten zu erheben. Die Nachnutzung von qualitativen Daten unterstützt auch größere Längsschnittforschung, indem Datensätze kombiniert werden, um mehr Teilnehmende zu analysieren. Gleichzeitig können qualitative Forschungsdaten zunehmend aus Online-Quellen erhoben werden. Sozialwissenschaftler*innen können über soziale Medien wie Blogs, Vlogs, Online-Foren sowie Beiträge und Interaktionen von Websites sozialer Netzwerke wie Facebook und Twitter auf persönliche Erzählungen und soziale Interaktionen zugreifen und diese analysieren. Diese großen sozialen Daten (big social data) wurden als beispiellose Quelle für Datenanalysen zelebriert, die in großem Umfang Erkenntnisse über das menschliche Verhalten liefern können. Diese Studie befasst sich mit den folgenden Forschungsfragen. **RQ1:** Wie unterscheidet sich die Kuratierung von Big Social Data von der Kuratierung qualitativer Daten? **RQ1a:** Wie unterscheiden oder ähneln sich epistemologische, ethische und rechtliche Fragen bei der Nachnutzung qualitativer Daten und bei Big Social Research? **RQ1b:** Wie können Datenkuratierungspraktiken wie Metadaten und Archivierung einige dieser epistemologischen und ethischen Probleme unterstützen und lösen? **RQ2:** Welche Auswirkungen haben diese Ähnlichkeiten und Unterschiede auf die Kuratierung großer sozialer Daten und die Kuratierung qualitativer Daten, und was können wir aus der Kombination dieser beiden Communities lernen? Meine Forschung verwendet ein sozialkonstruktivistisches Paradigma und befasst sich mit den Theorien der Communities of Practice und epistemischen Kulturen. Ich beantwortete meine Forschungsfragen durch eine eingehende Literaturanalyse und semi-strukturierte Interviews. Bei der Literaturanalyse habe ich sechs zentrale Gemeinsamkeiten zwischen der Nachnutzung qualitativer Daten und Big Social Research identifiziert. Die Interviews wurden mit drei unterschiedlichen Communities of Practices durchgeführt: qualitativ Forschende, Big Social Data Forschende und Datenkurator*innen. Ich habe die Critical-Incident-Technik verwendet, um den Interviewleitfaden zu strukturieren und die Methodik der Grounded Theory befolgt, um eine

qualitative Inhaltsanalyse der Interviewtranskripte durchzuführen. Diese Dissertation brachte folgende wichtige Erkenntnisse hervor. Erstens identifizierte diese Forschung sechs Schlüsselthemen für die qualitative Datennachnutzung und Big Social Data: Kontext, Datenqualität und Vertrauenswürdigkeit, Datenvergleichbarkeit, informierte Einwilligung, Datenschutz und Vertraulichkeit sowie geistiges Eigentum und Dateneigentum. Zweitens zeigte diese Forschung, dass jede Praxisgemeinschaft – qualitativ Forschende, Big Social Data Forschende und Datenkurator*innen – jedes der sechs Themen aus einer anderen Perspektive betrachtete und somit unterschiedliche Dimensionen jedes Themas priorisierte. Diese Variation der Perspektiven zeigt, dass die Verbindung dieser drei Praxisgemeinschaften ein breiteres Verständnis der Schlüsselfragen unterstützen und daher zu epistemologisch fundierteren, ethischeren und rechtlicheren Forschungspraktiken führen kann. Drittens stellt diese Dissertation fest, dass Datenkurator*innen gut positioniert sind, um Leitlinien für eine epistemologisch fundierte, ethische und rechtliche qualitative Datennachnutzung und Big Social Data bereitzustellen. Sie haben die Fähigkeiten und Perspektiven, um zwischen Praxisgemeinschaften zu übersetzen, und die Kompetenzen, sich um beide Arten von Daten zu kümmern.

Acknowledgements

Many thanks to the people who helped me complete this dissertation. To Vivien Petras for being a steadfast, kind, and patient advisor; I can't thank you enough for providing structure and guidance, and for instilling the process with a sense of delight. To Michael Zimmer for serving as my external advisor and providing key insights and feedback. To the students and mentors at the 2020 ASIS&T Doctoral Colloquium, especially Kalpana Shankar. To Eric Raile of the Montana State University (MSU) HELPS Lab for reviewing my interview guides. To Emily O'Brien for editing interview transcripts and to Jeanine Olson for reviewing my citations. I am also grateful to the 30 researchers and curators whose interviews were essential to this research; as one of them told me, qualitative research participants give the gift of their knowledge and experience to advance scientific discovery. I hope this research does justice to the gifts that they have given. To Kenning Arlitsch, Dean of the MSU Library, for supporting my pursuit of a doctoral degree, for connecting me with Vivien, and for sharing such well-organized and thorough materials relating to the processes at HU. To Venice Baird for conversations and collaborations that always produce new knowledge and insights into data curation and data sharing. To Doralyn Rossmann, Jodi Allison-Bunnell, and Jason Clark for supporting my research interests and providing me with guidance and care over the years as my department heads in the MSU Library. Jason, thank you for being my mentor and collaborating with me on so many exciting projects over the years; your people-centered values and your enthusiasm for our work have been transformative to my career.

Many thanks to my friends and family, and to my music and dance collaborators; you are constant sources of balance and happiness. To Scott Young, my partner and the other member of my doctoral cohort; our discussions, planning sessions, and mutual support were vital to my completion of this research. To my dad, David Mannheimer, for reading and editing my dissertation; it was so special to share my research with you and benefit from your keen eye for writing style, clarity, and structure. And to my sisters, Rachel Mannheimer and Katie Mannheimer, both of whom recently completed their own major writing projects; you inspire me and bring me so much joy! I wish Mom could be here to celebrate with us.

Table of contents

List of figures

List of tables

List of abbreviations

AoIR = Association of Internet Researchers

API = application programming interface

CAQDAS = Computer assisted qualitative data analysis software

CC0 = Creative Commons Public Domain Designation

CFAA = Computer Fraud and Abuse Act

DDI = Data Documentation Initiative

DocNow = Documenting the Now

DOI = Digital object identifier

FAIR = Findable, accessible, interoperable, and reusable

GDPR = European Union General Data Protection Regulations

HELPS Lab = The Human Ecology Learning and Problem Solving Lab

ICPSR = Inter-university Consortium for Political and Social Research

IP = intellectual property

IRB = institutional review board

MSU = Montana State University

OCAP = First Nations principles of ownership, control, access, and possession

OHRP = U.S. Health and Human Services Office of Human Research Protections

PI = Principal Investigator

QDAS = Qualitative data analysis software

QDR = Qualitative Data Repository

QuDEx = The Qualitative Data Exchange Schema

REFI = Rotterdam Exchange Format Initiative

SACHRP = U.S. Health and Human Services' Secretary's Advisory Committee on Human Research Protections

TEI = The Text Encoding Initiative

Chapter 1. Introduction

> Before social scientists can begin using ideas and algorithms from computer science, they need to learn how to work with large-scale unstructured organic data and understand the general principles, tools, and methods used by computer scientists. Likewise, computer scientists can reach inaccurate conclusions if they fail to understand key considerations and objectives within social science research that may not traditionally apply in computer science. (Mneimneh et al., 2021, p. 3)

1.1. Background

Recent years have seen the rise of exciting innovations in data sources and methods for social science research. My dissertation research was prompted by a desire to better understand the impact of these innovations on qualitative researchers and big social researchers, as well as on my own scholarly community of librarians and archivists who curate qualitative and big social data.

Trends toward open science practices, along with advances in technology, have promoted increased data archiving in recent years, thus bringing new attention to the reuse of archived qualitative data (Corti et al., 2005; Glenna et al., 2019). Qualitative data reuse has a variety of potential benefits, including increasing efficiency, deepening research conclusions, and reducing the burden on research subjects by allowing new studies to be conducted without collecting new data. Qualitative data reuse also supports larger-scale, longitudinal research by facilitating the combining of datasets to analyze more participants and to investigate human behavior over longer periods of time. In 2002, Mason encouraged the social science community to invest in longitudinal qualitative studies that were specifically designed for secondary use. She called for “appropriately qualitative ways to ‘scale up’ research resources currently generated through multiple small-scale studies, to fully exploit the massive potential that qualitative research offers for making cross-contextual generalisations” (Mason, 2002, as quoted in Davidson et al., 2018, p. 364). In the two decades since Mason issued this call, some researchers have aggregated qualitative data to

produce new conclusions (Davidson et al., 2018; Halford & Savage, 2017; Winskell et al., 2018), but it is still a rare practice.

At the same time, qualitative data can increasingly be collected from online sources. Researchers can access and analyze personal narratives and social interactions through social media such as blogs, vlogs, online forums, and posts and interactions on platforms like Facebook and Twitter. These "big social data" (Manovich, 2012) have been celebrated as an unprecedented source of data analytics, able to produce social insights by analyzing human behavior on a massive scale (Cappella, 2017; Fan & Gordon, 2014). Big social data is a form of qualitative data that has been published online by social media users themselves. When researchers analyze big social data, they could be considered to be reusing qualitative data—repurposing and recontextualizing this data to answer research questions.

Both the researchers who reuse qualitative data and big social researchers aim to scale up and enhance social science research. However, these two communities of practice are under-connected; big social research has not yet been widely framed as a form of qualitative data reuse, and qualitative data reuse has only begun to be discussed through a big social data lens. Additionally, these two communities of practice have different backgrounds, training, and disciplinary values. Big social researchers tend to have computer science and other types of engineering backgrounds, and they tend to focus on using computational methods to analyze large amounts of data. Qualitative researchers, on the other hand, tend to come from social science disciplines, and they tend to focus on using in-depth research methods to investigate social and behavioral phenomena.

1.2. Issues raised by qualitative data reuse and big social data

There are similarities and differences regarding risks and benefits when conducting research with big social research and qualitative data reuse. This dissertation is a comparative study of the communities of practice who conduct big social research and qualitative data reuse; my aim is to build a better understanding of how each community of practice could benefit from the other's practice. My dissertation also studies data curators as a third community of practice whose professional expertise and services can encourage and support responsible

social research and data sharing. Through in-depth literature reviews in Chapters 2 and 3, this dissertation identifies six key epistemological, ethical, and legal issues that are common to qualitative data reuse and big social data research. These key issues are context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. Chapters 2, 3, and 4 also discuss data curation strategies and initiatives that can alleviate some of these issues. The six issues are summarized below.

1.2.1. Context

Both archived qualitative data and big social data are context-dependent. For archived qualitative data, there is some concern that these data may not be able to be properly understood outside of their original context, without the knowledge and expertise of the researchers who conducted the original research project and originally analyzed the data (Hammersley, 2010; Walters, 2009). As Broom, Cheshire, and Emmison (2009) write, "the idea that data can be neutralized and deposited into an archive, ready to be 'picked up' by others, sits uncomfortably for many" (p. 1164). For big social data, context is even more murky—the context of a social media post may be absent or difficult to understand. Indeed, boyd and Crawford ask whether context and meaning can ever be accurately understood by big social data researchers (boyd & Crawford, 2012).

1.2.2. Data quality and trustworthiness

Issues relating to data quality and trustworthiness are also common to both big social research and qualitative data reuse. Researchers who reuse qualitative data need to know that the data they are using are high-quality and trustworthy—that the data have been collected using valid methods, that transcriptions are accurate, and that the data are complete. For big social data, social media users may not be representative of society as a whole, and the data collected through web scraping or calls to Application Programming Interfaces (APIs) may not be complete. Issues of trust are further complicated by the possibility of fake social media accounts and bots, which may appear to be human, but which researchers may not want to include a qualitative analysis of social media users.

1.2.3. Data comparability

For qualitative data reuse, the unstructured, complex, and varied nature of qualitative data can make it difficult to analyze a primary dataset so that it yields a meaningful answer to a secondary research question. Big social data may have different filetypes, different metadata fields, and different metadata standards, all of which make combining data more difficult, especially on a large scale. Data comparability is an important issue for both qualitative data reuse and big social research, because combining and comparing datasets helps enhance the context and quality of their research. Combining datasets can also increase the scope of qualitative and big social research by allowing researchers to build larger or longitudinal datasets.

1.2.4. Informed consent

Informed consent is an issue for both qualitative data reuse and big social research. For archived qualitative data, while research participants provide consent for the initial study, they may not have provided consent for the data to be archived for future use. In recent years, reuse clauses have begun to be written into consent documentation, and Institutional Review Boards (IRBs) can provide guidelines for consent procedures that allow the use of qualitative data beyond their original purpose (Elman et al., 2017). Big social researchers may not consider it necessary to obtain informed consent from the users who generate big social data, since they often consider big social data to be content that is simply found online. Big social researchers may also consider it sufficient that users have agreed to their social media platforms' terms of service; these terms generally include broad consent to data use, including research use. However, most users do not read the terms of service closely enough to constitute informed consent.

1.2.5. Privacy and confidentiality

Researchers who share and reuse qualitative data and big social researchers both contend with the issue of privacy and confidentiality. While some big social researchers have argued that big social data are public by nature, and therefore that deidentification of such data is unnecessary, negative public responses to projects such as the Taste, Ties, and Time dataset (Zimmer, 2010) and an openly shared OKCupid dataset (Resnick, 2016) have shown the perils

of sharing big social data without proper deidentification. For both qualitative and big social data, protecting participant privacy and confidentiality is all the more vital when participants are part of vulnerable populations such as prisoners, children, people involved in illegal activities, and marginalized and minoritized communities such as Black, Indigenous, LGBTQIA+, or disabled communities. Participants from these communities may face high risk if the deidentified data are able to be reidentified (Rothstein, 2010).

1.2.6. Intellectual property and data ownership

Intellectual property (IP) and data ownership is a key issue for both qualitative researchers who share or reuse data and big social researchers. Both communities of practice may encounter challenges when collecting existing data from sources where intellectual property rights, licenses, or permissions may be varied. For qualitative data, data may be owned by institutions, or intellectual property may be held by research participants. In either case, consent from IP rights holders is necessary to redistribute the data for reuse. For big social data, the IP rights are often controlled by private, for-profit companies. Even if social media posts are the intellectual property of the users who posted them, the rights to these posts are licensed to the social media companies through the companies' terms of service. Additionally, intellectual property rights and data ownership may vary according to how and where the data were collected. For example, when collecting data from Indigenous communities, additional considerations come into play, such as the CARE Principles (Carroll et al., 2021) and the First Nations principles of ownership, control, access, and possession (OCAP) (FNIGC, 2010).

1.3. Data curation to facilitate qualitative data reuse and big social research

The rapidly-evolving data landscape presents interesting possibilities for social and behavioral research. But as more researchers share data, more researchers also need help facilitating responsible research, data sharing, and data reuse practices. The field of data curation has grown exponentially in response to this need. However, data sharing practices and guidelines that are specific to qualitative data and big social data are still in the early stages of development. When confronting issues involving responsible data sharing and

reuse, data curators often refer to the FAIR Guiding Principles (Wilkinson et al., 2016), which suggest that shared data should be findable, accessible, interoperable, and reusable. However, the FAIR Principles were designed to support technical issues relating to data reuse. They do not directly address the ethical, epistemological, and legal issues that arise when using data originally created through interaction with human subjects.

A growing body of literature suggests that data curation can provide strategies to alleviate some of the issues described above. These practices include data management planning, promoting research design that facilitates later data sharing, and producing metadata and other documentation to capture contextual information (Elman et al., 2010; Thorne, 1994). Data curation can also provide strategies that help protect participants from harm, including data deidentification, amalgamating or aggregating data, and restricting access to data. (A. Clark, 2006; S. L. Garfinkel, 2015; Heaton, 2004). Qualitative data reuse is a more established practice, and literature going back to the 1990s explores how data curation strategies such as these can support epistemologically-sound, ethical, and legal data sharing. Data curation for big social data is less well-developed, and there is little consensus about how to maintain a balance between transparency and protecting research subjects.

1.4. Research questions and methods

My dissertation hypothesizes that comparing qualitative data reuse and big social research will support responsible research practices and improved data curation practices for both qualitative data reuse and big social research. By understanding the similarities and differences between qualitative data reuse and big social research, researchers and data curators can build stronger strategies for responsible use and reuse of qualitative data, both big and small. These strategies can reduce the potential for harm to the human subjects whose thoughts and activities are represented in archived qualitative data and big social data, while at the same time promoting the reuse of these datasets. My dissertation aims to answer the following research questions.

RQ1: How is big social data curation similar to and different from qualitative data curation?

RQ1a: How are epistemological, ethical, and legal issues different or similar for qualitative data reuse and big social research?

RQ1b: How can data curation practices such as metadata and archiving support and resolve some of these epistemological and ethical issues?

RQ2: What are the implications of these similarities and differences for big social data curation and qualitative data curation, and what can we learn from combining these two conversations?

My ultimate aim in conducting this research is to understand how the ideas and approaches of two distinct research communities—qualitative researchers and big social researchers—can be combined so that a third community—data curators—can develop and encourage stronger data curation practices, thus leading to more ethical, legal, and epistemologically sound qualitative data reuse and big social research.

Accomplishing this aim requires an in-depth understanding of researchers' behaviors and attitudes. Such in-depth understanding can be facilitated by a qualitative approach, rooted in grounded theory (Glaser & Strauss, 1967), to iteratively produce insights. My research followed a five-stage process, using the critical incident technique: (1) determine the general aims of the activity to be studied; (2) set specifications for data collection, including the types of situations to be observed or reported and the incident's relevance and effect on the general aim of the activity; (3) collect data via interviews or questionnaires centered around relevant incidents; (4) analyze the data; and (5) interpret and report the findings.

In Stage 1, *determine the general aims of the activity to be studied*, I reviewed the literature to identify the epistemological, ethical, and legal issues common to both qualitative data reuse and big social research. I used the methods outlined by Creswell (2009) and *The handbook of research synthesis and meta-analysis* (Cooper et al., 2019) to conduct a review of the literature on qualitative data reuse (Chapter 2) and big social research (Chapter 3). The data analysis focused on "analyzing whether and why there are differences in the outcomes of studies" (Cooper et al., 2019, p. 14). The last step of the literature review was

to interpret the literature. This step revealed six key epistemological, ethical, and legal issues—context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership—and identified central characteristics of these six issues to be investigated further.

In Stage 2 of the research process, I *set specifications for data collection, including the types of situations to be observed or reported and the incident's relevance and effect on the general aim of the activity.* In Stage 3 of the research process, I *collected data via interviews or questionnaires centered around relevant incidents*. I collected data using semi-structured interviews that centered around specific incidents of qualitative data archiving or reuse, big social research, or data curation. Semi-structured interviews have been used to study data sharing behaviors and attitudes (e.g., Faniel et al., 2019; Faniel & Connaway, 2018; Yoon, 2017; Zimmerman, 2008) and to study the behaviors and attitudes of communities of practice and epistemic cultures (Ardichvili et al., 2006; Keller & Poferl, 2016). This method is commonly used in grounded theory research because the researcher has "more direct control over the construction of data than does a researcher using most other methods, such as ethnography or textual analysis" (Charmaz, 2001, p. 676). The flexibility of open-ended questions gives researchers more analytic control over the data; as new ideas continually emerge throughout the interview process, the interviewer has the flexibility to pursue these new ideas (Charmaz, 2001). The key issues I had identified through my literature review informed this stage of my research. Potential interviewees were identified during the literature review process by contacting authors of data archived in repositories, and by contacting authors of relevant articles. After I identified an initial group of participants, I added participants using snowball sampling, an established method for augmenting a participant list, first developed in the 1960s (Kadushin, 1968). This sampling method is often used when interviewing potential participants who come from a relatively small professional population and who are therefore likely to be connected to each other (Bernard et al., 2017). Thus, snowball sampling is an appropriate method for this dissertation, which focuses on communities of practice who are conducting specialized research, data sharing, and curation activities. In addition to snowball sampling, I used theoretical sampling—that is, responsive sampling conducted at the same time as my interviewing and data analysis. By using theoretical sampling, I was able to selectively

identify potential participants according to the concepts I had derived from my analysis and any questions or gaps I identified along the way (Corbin & Strauss, 2008).

The semi-structured interviews were conducted as outlined in Luo and Wildemuth (2017). The "incidents" elicited from participants focused on one of five experiences, depending on each participant's community of practice: (1) big social research; (2) big social data archiving (3) qualitative data archiving (4) qualitative data reuse, or (5) big social or qualitative data curation.

I analyzed the interviews using a conventional qualitative content analysis of the interview transcripts. Because the interviews were structured around each of the six key issues that I identified in the literature review (see Chapters 2 and 3), I deductively created a parent code for each of the six key issues—context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. I then used inductive coding to create subcodes beneath each of the parent codes for these key issues. These approaches are outlined in Zhang and Wildemuth (2017) and detailed in Bernard, Wutich, and Ryan (2017). After coding each transcript, I normalized the themes by comparing any new themes to previous themes, in accordance with grounded theory's constant comparative method (Glaser & Strauss, 1967). Through this iterative process, I developed and documented coding rules that were applied to all of the interview transcripts. The interview analysis resulted in themes that aligned with the six key issues identified in the literature review; the analysis also produced three additional analytically-powerful themes.

The 30 interviews with big social researchers, qualitative researchers, and data curators demonstrated that the original six themes identified in my literature review were the appropriate categories with which to group the interviews. Each group of participants had clear ideas about, and responses to, each of these six themes. Additionally, my post-interview deductive coding process revealed three more themes: domain differences, strategies for responsible practice, and data curation issues. These three themes proved to be analytically powerful lenses through which the participants viewed big social research and qualitative data reuse—how each community of practice understood their own

disciplinary and methodological foundations and landscapes, the strategies that each community of practice used to support responsible practice, and each community of practice's experience with data curation.

The results of my research suggest that data curation strategies can support and enhance responsible practice in some cases, and that data curators can act as facilitators and intermediaries between communities of practice.

1.5. Structure of dissertation

This dissertation is structured as follows. Chapters 2 and 3 define key terms and review the literature. Chapter 2 defines qualitative data and qualitative data reuse, then identifies and discusses the six key issues that arise when sharing and reusing qualitative data—context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. Chapter 3 defines big social data and big social research, then discusses how the same six key issues apply when conducting big social research and sharing big social data. Chapter 4 provides a synthesis of data curation issues as they apply to the six common issues identified in Chapters 2 and 3. Establishing these six common issues is a key contribution of this dissertation, and these six issues help structure the interviews with researchers and curators that are described in more detail in Chapters 5 and 6.

Chapter 5 describes the theoretical framework underlying this research, and provides an in-depth review of my research methods. Chapter 6 supplies an in-depth review of the results of my semi-structured interviews with qualitative researchers, big social researchers, and data curators.

Chapter 7 discusses the results of my research and the implications for data curation. Chapter 8 concludes the dissertation with a discussion of key contributions, research limitations, and future directions.

1.6. Chapter summary

This research advances understanding of qualitative data reuse, big social research, and data curation. Big social researchers can conduct more epistemologically sound, ethical, and legal research by engaging with aspects of the six key issues that are important to qualitative data reuse. At the same time, qualitative research can be scaled up by engaging with aspects of the six key issues that are important to big social researchers. Ultimately, this dissertation proposes that data curators can apply the conclusions of this research to enhance their understanding of these two research communities, thus allowing data curators to provide a range of skills and services to support responsible big social research and qualitative data sharing. Data curators can provide researchers with tools, strategies, and guidance for epistemologically sound, ethical, and legal data sharing and reuse.

Additionally, this research suggests that data curators can also act as intermediaries between the communities of practice who conduct qualitative data reuse and big social research. By investigating how qualitative researchers, big social researchers, and data curators approach key issues, this dissertation aims to help data curators better position themselves to connect these other two communities of practice and to facilitate responsible qualitative data reuse and big social research.

Chapter 2. Literature review - Qualitative data reuse

In this chapter, I define qualitative data and qualitative data reuse, and I provide examples of types of qualitative data reuse. I then provide an overview of the history of qualitative data reuse and review the benefits of reusing qualitative data. Next, I discuss the challenges of data reuse, including epistemological, ethical, and legal issues. Lastly, I provide information about how data curation practices can support epistemologically sound, ethical, and legal qualitative data reuse.

2.1. Defining qualitative data and qualitative data reuse

2.1.1. Qualitative data

The focus of this chapter is qualitative data. In contrast to quantitative data, qualitative data are non-numeric (Kitchin, 2014, p. 5). These data may be analyzed to produce numeric results such as code counts and statistics, but the foundational qualitative data themselves are non-numeric (DuBois et al., 2018; Greener, 2011).

Bernard et al. (1986) define the "construction of primary data" in anthropology as "an interactive process between a researcher, a theory, and the research materials under study, whether they be people in the field or documents to be examined" (p. 363); Bernard et al. suggest four main types of data "construction:" "(1) relatively open-ended, unstructured interviews with key informants, (2) structured interviews of respondents who, in the case of surveys, may number in the hundreds or thousands, (3) direct observation of behavior and environmental features, and (4) extraction of information from existing records such as native texts, court proceedings, marriage records, and so on" (p. 382). As these passages suggest, qualitative data are produced by qualitative research, and the data can be defined by the process that was used for creating or collecting them. The National Endowment for the Humanities Office of Digital Humanities corroborates this idea, defining data as "materials generated or collected during the course of conducting research" (2019, p. 1). Corti describes qualitative research as "defined by openness and inclusiveness, aiming to capture participants' lived experiences of the world and the meanings they attach to these experiences from their own perspectives" (Corti, 1999, p. 19). In order to meet the aims

described by Corti, qualitative researchers collect and examine various types of data. Bernard, Wutich, and Ryan (Bernard et al., 2017) suggest that qualitative data exist in five formats: (1) physical objects, (2) still images, (3), sounds, (4) moving images; and (5) texts. Their table, *Kinds of qualitative data based on form, size, and accessibility*, is adapted below as Table 1.

Table 1. Kinds of qualitative data based on form, size, and accessibility (adapted from Bernard et al., 2017, p. 11)

	Small		**Large**	
	Public	**Private**	**Public**	**Private**
Physical Objects	Park sculptures, street signs, pottery shards, store merchandise	Personal jewelry, pill bottles, blood samples	Archaeological ruins, buildings, houses, universities, skyscrapers	Household garbage, clothing
Still Images	Magazine ads, cave art, billboards, webpages, paintings hung in museums	Doodles, line sketches, family portraits (analog or digital), patient x-rays	Large detailed murals, art exhibits (analog or digital)	Family albums (analog or digital), art portfolios (analog or digital), CAT scans
Sounds	Jingles, radio ads, intercom announcements, messages you hear while on hold	Memo dictation, answering machine messages, elevator conversations	Political speeches, sports play-by-plays, music albums, focus group recordings	Oral histories, demo soundtracks, in-depth conversations, clinical interviews
Moving Images: Video	TV ads, news footage, sitcoms, TikTok videos	Home-movie clips (analog or digital)	Full-length movies, documentaries, television programs	Long video recordings of events like family reunions and weddings

Texts	Epitaphs, obituaries, personal ads, parking tickets, Twitter posts using hashtags	Thank-you letters, shopping lists, short responses to interview questions, emails, diaries	Books, manuals, court transcripts, congressional record and data, newspapers, news websites	Diaries, detailed correspondence, private online forum discussions

The types of data identified in Table 1 are far-reaching, and include many types of artifacts and objects that a qualitative researcher could analyze. Bernard et al. do not specify the format of the qualitative data examples in their book, but these types of data can include born-digital or digitized data such as digital photographs, digital voice memos, digital video recordings, ebooks, and word processing documents. To make clear that qualitative data can be both analog and digital, I have updated Bernard et al.'s table to include a few specific examples of digital qualitative data.

Heaton provides a simplified classification structure for qualitative data, dividing these different formats into "non-naturalistic" data (data that are solicited by researchers through interviews, questionnaires, etc.), and "naturalistic" data (data that are found or collected by researchers with minimal interaction with the research subjects) (see Table 2).
Non-naturalistic qualitative data may take the form of fieldnotes and other observational records of solicited interactions with the research subjects, interviews, focus groups, solicited narratives, and questionnaires with open-ended questions. Naturalistic qualitative data may take the form of autobiographies, found narratives, letters, official documents, photographs and film, and observation of social interactions made without intervention from the researchers (Heaton, 2004).

Table 2. Examples of qualitative data used in social research (adapted from Heaton, 2004, p. 15)

Type	Examples
Non-naturalistic or artifactual data (solicited for research studies)	Fieldnotes Observational records Interviews Focus groups Questionnaires (responses to open-ended questions) Diaries (solicited) Life stories
Naturalistic data (found or collected with minimal interference by researchers)	Life stories Autobiographies Diaries (found) Letters Official documents Photographs Film Social interaction

As in Bernard et al., Heaton does not specify the format of the data in Table 2. Each of these types of data listed in Table 2 could be either analog or digital. For example, fieldnotes could take the form of paper notebooks or word processing documents; diaries could be written using pen and paper, kept using the iPhone Notes app, or openly posted online in blog form; and social interactions could take the form of a face-to-face conversation or a technology-mediated interaction such as a Twitter exchange or a Reddit thread.

For purposes of this dissertation, taking into account the kinds of data listed in Tables 1 and 2, I define qualitative data as analog or digital objects, images, sounds, moving images, and texts that are collected and/or analyzed by researchers during the course of qualitative research.

2.1.2. Qualitative data reuse

The term “secondary analysis” has been used since the mid-20th century to describe a research methodology that uses pre-existing data (whether quantitative or qualitative). Lipset and Bendix provide a simple definition of this concept: “the study of specific problems through analysis of existing data which were originally collected for another purpose” (Lipset

& Bendix, 1959, p. ix). It should be noted that secondary analysis is distinct from meta-analysis and literature review. Meta-analysis and literature review synthesize research findings, whereas secondary analysis uses primary data to generate new insights (Heaton, 1998; Thorne, 1998)

The definitions of secondary analysis developed over the decades clarify this distinction. For instance, Glass suggests that secondary analysis is conducted for the purpose of "answering the original research question with better statistical techniques or answering new questions with old data" (Glass, 1976, p. 3), and Hakim defines secondary analysis as "further analysis of an existing data set which presents interpretations, conclusions, or knowledge additional to, or different from, those presented in the first report on the enquiry as a whole and its main results" (Hakim, 1982, p. 2). In her 2004 definition of qualitative secondary analysis, Heaton additionally brings in the idea of verification, writing that "secondary analysis is a research strategy which makes use of … preexisting qualitative research data for the purposes of investigating new questions *or verifying previous studies* [emphasis added]" (Heaton, 2004, p. 24). In order to explain this definition, it is necessary to discuss the concept of verification in qualitative research.

In the 1970s and 1980s, verification was considered a way to legitimize qualitative research—to prove its dependability, confirmability, and trustworthiness (Guba, 1981; Guba & Lincoln, 1989; Heaton, 2004; Scheff, 1986). However, as discussion of qualitative data sharing increased in the 1990s and 2000s, some began to argue that verification might not be applicable to qualitative research—suggesting that the phenomena studied by qualitative researchers are too heterogeneous to be verified or audited. As Hammersley writes in 1997, "these phenomena are locally distinctive, changing in character both over time and across social contexts, and data about them are subject to reactivity, to distortion arising from the research process itself. The potential for replication in any strict sense is therefore quite limited" (Hammersley, 1997, p. 132). Others argue that the auditing of qualitative data could "expose researchers to scrutiny which is counterproductive to both the institution of research and the interests of individuals involved" (Parry & Mauthner, 2004, p. 146). Corti suggests that "certain approaches used in qualitative research, for example, grounded theory which opposes the scientific paradigm of testing hypotheses, do not lend themselves

to verification" (Corti, 2000, section 6.1.). Stenbacka also argues that the overall concepts of validity and replicability are not generally applicable to qualitative research (Stenbacka, 2001).

Most recently, Tsai et al. declare verification to be difficult for qualitative research, due to "the inherently intersubjective nature of qualitative data collection, the iterative nature of qualitative data analysis, and the unique importance of interpretation as part of the core contribution of qualitative work" (2016, p. 192). Heaton suggests that, "in practice the closest qualitative researchers have traditionally come to verifying studies is through conducting additional primary research designed to emulate the original" (2004, p. 30).

Overall, while it may be rare or difficult to use qualitative data for verification purposes, such use of the data is theoretically possible. This possibility suggests that one should not completely exclude verification from the definition of secondary analysis. Nevertheless, in this dissertation I have opted not to use the term "verify" in my definition of secondary research; instead, I use the phrase "refine ideas" to reflect the concept that qualitative data can be used to review and refine previous research.

As demonstrated by the discussion above, a definition such as Thorne's—"the reexamination of one or more existing qualitatively derived data sets in order to pursue research questions *that are distinct from* [emphasis added] those of the original inquiries" (2004, para. 1)—may be too narrow. Qualitative data may be used to ask the same questions that were asked in the original research, but for different purposes. Qualitative data are often the result of participatory research—a co-creation process between researchers and participants, through observation and conversation. When researchers use archived qualitative data, they repurpose what were previously co-created data, introducing new contexts, potentially asking new research questions, and potentially gathering new data to augment the archived data. To reflect these ideas, Moore suggests that the ways in which qualitative data are reused can sometimes go beyond the traditional definition of "secondary analysis," so she reframes the practice as a "recontextualization" of data (Moore, 2007). Moore's idea of recontextualization aligns with current terminology. As data sharing and data publication become more common practices (see section 2.2.), the more recent focus is not necessarily

on secondary analysis as a methodology, but rather on the idea of data reuse to support research of many different types. Scholars have therefore begun to increasingly use the broader term "data reuse." Bishop and Kuula-Luumi suggest that "reuse provides an opportunity to study the raw materials of past research projects to gain methodological and substantive insights" (2017, p. 1). van de Sandt et al. take a broad view of data reuse, concluding that reuse can be seen as equal to use. They define reuse as "the use of any research resource regardless of when it is used, the purpose, the characteristics of the data and its user" (van de Sandt et al., 2019, Discussion section).

This dissertation adopts the broader term *qualitative data reuse*, using the following definition: Qualitative data reuse is when researchers use existing qualitative data to refine ideas, gain new insights, and produce new scholarship. (This dissertation limits its definition to the scholarly use of data.)

2.1.2.1. Types of qualitative data reuse

Several types of research involving qualitative data reuse have been defined in the literature. Thorne suggests five approaches: (1) "analytic expansion" in which a researcher reuses their own data in order to conduct further inquiry; (2) "retrospective interpretation" which expands upon questions that were raised in the original study, but were not central to that study; (3) "armchair induction" in which researchers use textual analysis to develop new theories; (4) "amplified sampling" in which several datasets are compared in order to establish broader theories; and (5) "cross-validation" in which existing data are used to validate new findings or show new patterns beyond the scope of individual research studies (Thorne, 1994, pp. 266–267).

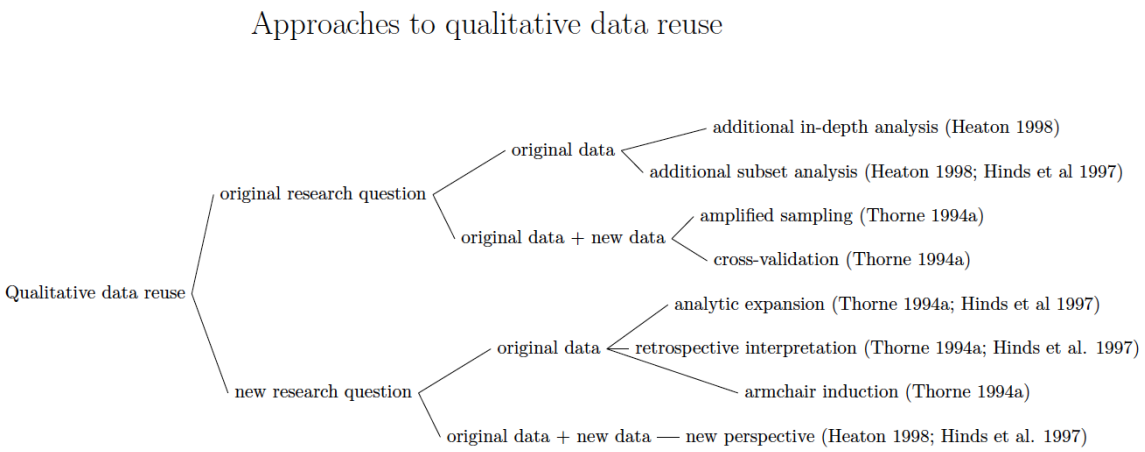
Hinds, Vogel, and Clarke-Steffen outline four approaches to qualitative analysis using existing data: (1) conducting new types of analyses that are different from those used in the original study; (2) analyzing a subset of the data for a similar, but more focused research study; (3) reanalyzing data by focusing on concepts that were not specifically addressed in the primary analysis; and (4) integrating existing qualitative data into a new study that refines the

purpose, questions, and data collection processes of the original study (Hinds et al., 1997, pp. 409–410).

Heaton outlines three approaches: (1) “additional in-depth analysis,” which places a more intensive focus on a particular part of the original study; (2) “additional sub-set analysis,” in which further analysis is conducted on a selected sub-set of the original data; and (3) “new perspective/conceptual focus,” in which a different perspective is applied to all or part of a data set, and the research reusing the data examines concepts that were not central to the original research (Heaton, 1998, What is Secondary Analysis? section).

In 2004, Heaton describes six different types of qualitative data reuse: “(1) Supra analysis: transcends the focus of the primary study from which the data were derived, examining new empirical, theoretical or methodological questions. (2) Amplified analysis: combines data from two or more primary studies for purposes of comparison or enlarging sample. (3) Supplementary analysis: a more in-depth investigation of an emergent issue or aspect of the data which was not addressed in the primary study. (4) Complementary analysis: the [secondary analysis] is supported by additional primary research or, alternatively, a primary study which includes an element of [secondary analysis]. (5) Alternative analysis: data are re-analysed using new methods and/or perspectives for purposes of corroboration based on the principle of triangulation. (6) Repeat analysis: data are re-analysed using a similar analytical framework in order to verify the findings of the primary research” (Heaton, 2004, p. 38). Figure 1 synthesizes the different approaches described above.

Figure 1. Approaches to qualitative data reuse (Flowchart inspired by Schöch, 2017; and van de Sandt et al., 2019)



The definitions above do not differentiate between data collected oneself or data collected by another researcher. While some suggest that reusing one's own data could reduce challenges and increase benefits (Heaton, 2004; Hinds et al., 1997; Sherif, 2018; Thorne, 1998), Mauthner, Parry, & Backett-Milburn (1998) write about the challenges they faced when revisiting their own data for analysis, suggesting that the passage of time caused reuse of even their own data to be difficult. Irwin (2013) argues that reusing one's own data provides a critical distance from which researchers can evaluate the quality and efficiency of the data from the perspective of new research questions, and they can identify and provide any missing information. Thus, this dissertation considers that all data reuse has similar benefits and challenges, regardless of who originally collected it. Whatever method is used while reusing existing data, the epistemological, ethical, and legal issues remain the same from a data curation perspective.

2.2. History and benefits of qualitative data reuse

The practice of data reuse goes back to the first part of the 20th century, when researchers began reusing survey data in an effort to "save time, money, careers, degrees, research interest, vitality, and talent, self-images and myriads of data from untimely, unnecessary, and unfortunate loss" (Glaser, 1963, p. 14). The earliest book describing secondary analysis in detail was published in 1972 (Hyman, 1972), and a major symposium, *Secondary Analysis of*

Existing Data Sets: For What Purpose and Under What Condition, was held at the Annual Meeting of the American Educational Research Association in New York in 1977. Since then, *quantitative data* reuse has generated an expansive body of literature, including educational texts on finding and analyzing statistical datasets (e.g., Hakim, 1982; E. Smith, 2008; Kiecolt & Nathan, 1985), and other literature examining the epistemological, ethical, and legal implications of reusing existing quantitative data in the social sciences (de Lusignan et al., 2007; Duke & Porter, 2013; Goodwin, 2012; Hartter et al., 2013).

As early as 1962, Glaser wrote that "secondary analysis is not limited to quantitative data. Observation notes, unstructured interviews, and documents can also be usefully reanalyzed. In fact, some field workers may be delighted to have their notes, long buried in their files, reanalyzed from another point of view" (Glaser, 1962, p. 74). However, despite this early mention, qualitative data reuse did not become a common practice until the 1990s (e.g., Corti, 1999; Hammersley, 1997; Heaton, 1998; Hinds et al., 1997; Mauthner et al., 1998; Szabo & Strang, 1997; Thompson, 2000; Thorne, 1994).

The practice of qualitative data reuse continued to grow through the 1990s and 2000s. Some still questioned whether reusing qualitative data was "tenable, given that it is often thought to involve an intersubjective relationship between the researcher and the researched" (Heaton, 1998, Methodological and Ethical Considerations section), but a growing faction of researchers, funding agencies, and academic journals began to increasingly consider data—both qualitative and quantitative—to be a public resource that should be formally published in addition to associated publications, especially for government-funded research (Dunn & Austin, 1998; Heaton, 2004)**.** The National Institutes of Health (NIH) began to require data sharing plans in its grant proposals in 2003, and recently released updated guidelines that will go into effect in 2023 (National Institutes of Health, 2020); the National Science Foundation introduced a data management plan requirement to support data sharing and reuse in 2011 (National Science Foundation, 2011); and the White House Office of Science and Technology Policy released a memo calling for a national commitment to data sharing in 2013 (Holdren, 2013). Private funders such as Wellcome (2017) and Gates Foundation (2015) have followed suit. Academic societies and journals have also adopted data sharing guidelines and policies. Examples include the American Psychological

Association (APA Data Sharing Working Group, 2015), the American Sociological Association (ASA, 2018, p. 16)*, American Economic Review* (Bernanke, 2004), *Journal of the Medical Library Association* (Akers et al., 2019), the Joint Data Archiving Policy (Dryad Digital Repository, 2011), and others (PLOS, 2014; Taichman et al., 2017). While the guidelines and policies outlined here are not specific to qualitative data, they have impacted the data sharing landscape, constituting a strong trend in the scientific community as a whole to encourage data sharing for the purpose of reuse.

Data sharing for qualitative data reuse was initially facilitated either by reusing one's own previously collected data, or through informal sharing between researchers (Heaton, 2008). However, more formal qualitative data sharing was bolstered by the creation of the United Kingdom's Qualidata, a social science qualitative data archive that aimed to curate and make available qualitative data on a national scale. Qualidata was launched in October 1994 (Corti & Thompson, 1996, 1998), and it was integrated into the UK Data Archive in the early 2000s. Since then, qualitative data archives have continued to be established. Examples in the United States include the Murray Research Archive at Harvard (Corti & Backhouse, 2005) and the Qualitative Data Repository, housed at the Center for Qualitative and Multi-Method Inquiry, a unit of Syracuse University's Maxwell School of Citizenship and Public Affairs (Elman et al., 2010; Karcher et al., 2016).

2.2.1. Benefits of qualitative data reuse

Qualitative data reuse has become more common in the 21st century as the scholarly community becomes more attuned to its potential benefits. As Mauthner writes, "the case for sharing data rests on three central pillars: a scientific, a moral, and an economic one" (Mauthner, 2012, p. 157).

The scientific benefits of qualitative data sharing include:

- Building new knowledge, new hypotheses, new methodologies, comparative research, and critiquing or strengthening existing theories (Corti, 2000; DuBois et al., 2018; Heaton, 1998; Jones et al., 2018). For example, the research dataset from the Timescapes Study, which explored how personal and family relationships developed

and changed over a 5-year period, has been used extensively by secondary researchers (DuBois et al., 2018).

- Promoting interdisciplinary use of data (Heaton, 2004; White, 1991). For example, the Human Relations Area Files (Murdock, 1961) are cultural materials from the field of anthropology that have been used to facilitate hypothesis-testing quantitative analyses (Ember, 2007), and have also been used for qualitative analysis, such as an exploratory analysis of household responses to water scarcity (Wutich & Brewis, 2014).
- Providing data for teaching purposes (Corti, 2000; Heaton, 1998; Jones et al., 2018; Sieber, 1991a; Szabo & Strang, 1997). For example, Bishop describes classroom assignments that faculty at universities in the United Kingdom have developed using data from the Qualidata repository to explore and evaluate qualitative research methods (Bishop, 2012).

The moral benefits include:

- Facilitating more research about rare, hard-to-reach, or inaccessible respondents while reducing the burden on research subjects (Heaton, 1998, 2004; Jones et al., 2018; Szabo & Strang, 1997). For example, Jones and Alexander (2018) describe how, during an oil and gas boom in the Canadian Arctic in the 1960s and 1970s, social scientists were increasingly interested in studying the effects of natural resource extraction on the four main indigenous communities in the area. Community members responded with concern about the number of studies being conducted, questioning whether the burden on participants yielded a corresponding benefit to their communities. Increased sharing of qualitative data supports new research without collecting new data and placing undue burden on communities who participate in the research.
- Transparency and accountability—in order to foster trust with the public and other researchers, and to share the results of public research funding (DuBois et al., 2018). This benefit is illustrated by the proliferation of data sharing policies among research funders, including NIH and NSF (see page 21).

Economic benefits include:

- Avoiding duplication of effort and allowing the conservation of time and resources, therefore supporting a higher return on investment (Fienberg et al., 1985; Hinds et al., 1997; Jones et al., 2018; Szabo & Strang, 1997; White, 1991). A 2013 study conducted on the UK's Economic and Social Data Service, Archaeology Data Service, and British Atmospheric Data Centre emphasized the economic benefit of data sharing, finding that: "very significant increases in research, teaching and studying efficiency were realised by the users as a result of their use of the data centres; the value to users exceeds the investment made in data sharing and curation via the centres in all three cases; and by facilitating additional use, the data centres significantly increase the measurable returns on investment in the creation/collection of the data hosted" (Beagrie & Houghton, 2014, pp. 4-5).

2.3. Issues in qualitative data reuse

Despite these potential benefits, qualitative data reuse raises a number of epistemological, ethical, and legal issues, which I will discuss further below.

2.3.1. Epistemological issues

Epistemological challenges in qualitative data reuse relate to context, data quality, and data comparability.

2.3.1.1. Context

Qualitative research is a process that may include deep and prolonged contact and connection with research subjects with the goal of understanding the subjects within their own context (Miles et al., 2020). Qualitative data are therefore highly context-dependent. As Hinds et al. write, "context is a source of data, meaning, and understanding… Ignoring context, underusing it, or not recognizing one's own context-driven perspective will result in incomplete or missed meaning and a misunderstanding of human phenomena" (Hinds et al., 1992, p. 72). The literature reflects the importance of considering whether data can be properly understood outside of their original context, without the nuanced knowledge and

expertise of the researchers who conducted the original research project and originally analyzed the data (Corti, 1999, 2000; Corti & Thompson, 1998; N. Fielding & Fielding, 2000; Hammersley, 1997, 2010; Hinds et al., 1997; Thompson, 2000; Thorne, 1994; Walters, 2009). As Broom, Cheshire, and Emmison write, "the idea that data can be neutralized and deposited into an archive, ready to be 'picked up' by others, sits uncomfortably for many" (2009, p. 1164). Dale, Arber, & Procter suggest that "it seems unlikely that the re-analysis of either interview transcripts or field notes by an outsider could give more than a partial understanding of the research issues" (1988, p. 15). Pasquetto, Borgman, and Wofford write that "removing data from their original context necessarily involves information loss" (2019, p. 23) stemming from small adjustments that may be made to the data during research and the loss of other deep knowledge of the research that data creators hold but may not be able to communicate in a dataset description; Pasquetto et al. suggest that collaboration with the original data creators can provide mutual benefit and support clearer contextual understanding. However, Mauthner and Parry discuss in several articles the difficulty of understanding context when reusing data, even when attempting to reuse data that they themselves had previously collected (Mauthner et al., 1998; Mauthner & Parry, 2009; Parry & Mauthner, 2004). Mauthner, Parry, and their coauthors suggest that insights are created through not only reviewing the data, but also through a deep knowledge of the research context and research subjects—that in qualitative research, "meaning is made rather than found" (Mauthner et al., 1998, p. 735). That meaning is made through the data collection process itself—which can be deeply affected by researchers' own cultural experiences, biases, and decision-making processes; it is additionally made through the process of data analysis, which is likewise affected by the unique perspective of the data analyst (Thorne, 1994; Tsai et al., 2016).

Some literature suggests that context is a challenge regardless of whether researchers are conducting primary or secondary research. Fielding argues that the challenge of context is less epistemological than practical, writing,

> "information regarded as vital in providing evidence for a given analytic point may well be missing from the archived data. But that happens in primary data analysis too—the tape runs out 'just when things get interesting', or the respondent withdraws their remark, or the observer leaves the police station just before the

> suspect gets violent, or any number of other contingencies. One might, and should, expect the professional researcher to respond to such a contingency in exactly the same way regardless of whether the data source is primary or secondary—by saying 'that is too bad but I cannot evidence this point' and moving on to what can be evidenced by the material available. Since one of the attractions of qualitative research is the richness of the data it can produce, this is not such a terrible problem" (2004, p. 99).

Regardless, there is some contextual information that can never be communicated. Thorne describes how field researchers "make mental notes of the conditions that make a single key informant more vehement, analytical, or articulate than the rest, features of the setting that might shape a particular instance of data, and an infinite number of details that influence direction but that may never become accessible within formal field notes" (Thorne, 1994, p. 268). Responding to the idea that some contextual information is either undocumented or undocumentable, some go so far as to say that data reusers should contact the researchers who originally collected the data (Heaton, 2008; Hinds et al., 1997; Szabo & Strang, 1997). However, this strategy is impractical for long-term use of data beyond the lifetime of the original researchers.

Hinds et al. frame distance from the original context of the data as a possible benefit, arguing that distance can free a researcher from developing fixed ideas about the phenomena reflected in the dataset, so long as the secondary researcher has enough knowledge of the original context to prevent misinterpretation (1997). Data curation strategies can also support communication of context. A number of scholars argue that contextual knowledge can be provided through proper metadata and documentation (Bernard et al., 2017; Corti, 1999, 2000; Elman & Kapiszewski, 2014; N. Fielding, 2004; Goodwin & O'Connor, 2006; Mannheimer et al., 2019; van den Berg, 2005). Bernard et al. urge primary researchers to be sure to accurately document both "the research procedures used and the social context" (1986, p. 384). Metadata and documentation are discussed in more detail in section 2.4.1.

2.3.1.2. Data quality and trust

Any reuse of qualitative data relies on the data's quality, especially when the data were collected by other researchers. Before the data can be reused, researchers need to spend time reviewing the dataset in order to assess the quality of the data (McCall & Appelbaum, 1991; Yoon, 2017). Thorne advises that "overt adherence to such dimensions as credibility, transferability, confirmability, and dependability creates the major mechanisms by which the trustworthiness or 'truth value' of the products of qualitative research can be evaluated" by the secondary data analyst (1994, pp. 274–275). Stenbacka suggests that the concepts of "validity, reliability, generalizability and carefulness… have grown to be generally accepted as having to be solved in order to claim a study as part of proper research" (2001, p. 551). Sherif advises that "the original data must allow the researcher conducting secondary analysis to understand examined processes, relationships, and subjective meanings" (Sherif, 2018, section 4, para. 3). I further examine the dimensions of data quality below.

Credibility can be understood through examining the credentials of the data creators and understanding other factors that affect the data collection such as training and time spent collecting data. (Hinds et al., 1997). Transferability/generalizability can be measured partly by examining the breadth and depth of the dataset, to determine whether the data are appropriate for reuse. Hinds et al. suggest reviewing three randomly selected interviews in order to determine whether the larger dataset can be used to achieve the research goals of any contemplated new study (1997). Data quality also relies on completeness and accuracy of the dataset.

Even if data are collected with care, there are multiple ways in which errors can be introduced that reduce the dependability of the dataset. Research subjects, reporters or recorders of field data, researchers, and data coders can all introduce errors. Simple mistakes or inaccuracies can occur throughout the process. Systematic errors can also be introduced into datasets as a result of bias related to personal identity, political ideology, general personality, or faculty assumptions. Bernard et al. suggest that "researchers using archival material need actively to consider potential biases and then, whenever possible, test for them" (1986, p. 391).

Data curators can contribute to data trustworthiness by co-producing data with data producers—providing data management, curation, and metadata support to increase data quality (Frank et al., 2017; Giarlo, 2013; ICPSR, 2019; Mannheimer et al., 2019; Yoon, 2017; Yoon & Lee, 2019). Data repositories and academic libraries also support trust through certifications such as the CoreTrustSeal Trustworthy Data Repositories Requirements (CoreTrustSeal, 2020) and the TRUST principles for digital repositories (Lin et al., 2020). I further discuss metadata and data archiving in section 2.4.

2.3.1.3. Data comparability

When reusing data, researchers must determine whether the primary data can be understood or analyzed in a way that is applicable to the study reusing the data—also referred to as data "fit." Because "qualitative research tends to produce data sets that are relatively unstructured, rich and diversified" (Heaton, 2004, p. 58), it can be difficult to fit a primary dataset into a secondary research question. Bernard et al. suggest that in some forms of qualitative research, such as unstructured interviewing, data may not be comparable across all informants (Bernard et al., 1986). However, Glaser suggests that comparability is possible, and that researchers should consider comparability across five dimensions: "1. populations, 2. situational dynamics, 3. problems under study, 4. variables or concepts, and 5. past findings with present hypotheses" (Glaser, 1962, p. 71). Generally, the literature suggests that comparability or "fit" can be determined using three strategies: (1) identifying the extent of missing data; (2) identifying how well the research questions converge in the primary research and secondary research; and (3) assessing the methods used to produce the primary data (Heaton, 2004; Hinds et al., 1997; Thorne, 1994). Another challenge for data comparability is that qualitative researchers often use proprietary qualitative data analysis software such as NVivo and Atlas.ti. These proprietary softwares may not be interoperable, and could cause challenges for data reuse. Some research has begun to support standardized formats and interoperability (Corti & Gregory, 2011; J. Evers et al., 2020), but more advocacy for this approach is needed. Data curators can support comparability of qualitative datasets by encouraging researchers who publish qualitative data to include clear documentation addressing missing data, research questions, and

methods, by using and encouraging standardized metadata, and by advocating for open source software and interoperable formats. Data curation strategies are further discussed in section 2.4.

2.3.2. Ethical and legal issues

In addition to laws and regulations, academics are guided by ethical frameworks. These ethical frameworks are built upon the values of the academic discipline and the guidelines of professional organizations and learned societies, as well as ethics regulatory guidance like the Nuremberg Code (BMJ, 1996), the Declaration of Helsinki (World Medical Association, 2013), the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979), and the Federal Policy for the Protection of Human Subjects, or "Common Rule" (U.S. Department of Health and Human Services, 1991). Most recently, the General Data Protection Regulations in the European Union have brought an increased awareness to ethical data use (Voigt & von dem Bussche, 2017). Ethical and legal challenges in qualitative data reuse relate to informed consent, confidentiality, and the intellectual property rights of research participants.

2.3.2.1. Informed consent

Qualitative researchers "have been involved in a long-standing debate about whether or not consent can ever be truly informed" due to the developmental, reflexive nature of research. (Parry & Mauthner, 2004, p. 146). And in fact, some go so far as to suggest implementing "process consent"—a structure in which research subjects continually consent to their participation as the researchers' ideas and inquiries evolve (Lawton, 2001). However, other researchers advocate for striking a balance that protects participants without overly obstructing the research process (Alexander et al., 2020; Wiles et al., 2007).

Consent for research involving qualitative data reuse is even more thorny. When reusing data from previous studies, some argue that consent should be re-obtained from the original participants. This strategy is also called the selective, repeated, or reconsent model, in which participants consent anew to each future use of their data (Joly et al., 2015; Master & Resnik, 2013). As Thorne writes, "there may be especially sensitive instances in which the

implied consent of original subjects cannot be presumed" (1994, p. 269). However, Heaton suggests that reconsent may be too difficult to be used often: "given that it is usually not feasible to seek additional consent, a professional judgement may have to be made about whether reuse of the data violates the contract made between subjects and the primary researchers" (1998, Methodological and Ethical Considerations section, number 4. Ethical Issues). Heaton later suggests that "it may be inappropriate to generalise about the need to obtain informed consent for secondary analyses, as this is likely to vary according to the characteristics of the secondary study" (Heaton, 2000, p. 3).

Heaton writes, "as part of the process of obtaining consent for participation in primary studies, research participants should be informed about any possibility that the information they provide may be shared with others" (Heaton, 2004, p. 78). Hinds et al. reiterate this idea, writing that "a researcher planning a secondary analysis will doubtlessly feel more ethically correct if permission from the participants in the primary study has been solicited at the time of the primary study" (1997, p. 414). Tiered consent (also called flexible consent, line-item consent, or multilayered consent) is a common strategy to provide a wider variety of consent options for participants, and can be useful for research in which participants consent to data reuse. Tiered consent provides participants with options for data sharing—opting out completely, consenting to restricted data sharing only, allowing participants the opportunity to review the data prior to sharing, and other options (Tiffin, 2018; VandeVusse et al., 2022). Regardless of consent strategy, questions remain about how well research participants understand the full implications of data sharing. In a recent study on abortion reporting, VandeVusse, Mueller, and Karcher's found that many participants who agreed to "data sharing" misunderstood the term to mean dissemination of research results, even through the consent form contained a detailed description of how the research data would be shared (VandeVusse et al., 2022).

The General Data Protection Guidelines (GDPR) in the European Union regulate and define the obligation to communicate clearly about data sharing. The GDPR requires that if a data controller (i.e., a person or organization that controls data processing) "intends to process personal data for a purpose other than that for which it was collected, it should provide the data subject prior to that further processing with information on that other purpose and

other necessary information" (Voigt & von dem Bussche, 2017, p. 17). A comparable set of guidelines does not exist in the United States.[1] However, the revised Common Rule, which went into effect in 2019, adds more explicit guidelines for secondary research, including the idea of broad consent—that is, when participants' consent includes "future storage, maintenance, or research uses" of their data (U.S. Department of Health and Human Services, 2017, Recommendations on the Interpretation and Implementation of Broad Consent section, para. 1). While secondary data use is still viewed as exempt from review, Exemption 7 and Exemption 8 in the revised Common Rule now explicitly state that broad consent must be obtained from primary research participants in order for secondary research with identifiable human subjects data to be considered exempt (Office for Human Research Protections, 2018). Institutional Review Boards (IRBs) that oversee ethical practice in human subjects research in accordance with the Common Rule are increasingly beginning to provide specific language that researchers can use to obtain broad consent and thus support data reuse (Cornell Research Services, 2019; Elman et al., 2018; Lavori et al., 1999; Siminoff, 2003), and in July 2021, NIH released a request for information about developing consent language for data reuse, indicating that such language may increasingly be standardized (Office of The Director, 2021).

However, broad consent is not a perfect solution, especially when viewed through the lens of feminist and post-colonial theories, which consider power structures between researchers and research subjects. There is concern that broad consent "exposes respondents to risk and uncertainty … [and] marginalizes respondents' moral and political rights to retain on-going involvement and decision-making powers in how their data will be used in the future" (Mauthner & Parry, 2013, p. 60).

2.3.2.2. Privacy and confidentiality

When sharing qualitative data for future reuse, researchers use various strategies to protect the confidentiality of participants in adherence to ethical and legal standards. Data deidentification procedures attempt to disguise the identity of participants by deleting their

[1] The California Consumer Privacy Act, which went into effect in the state of California in January 2020, dictates that "a business that sells the personal information of consumers shall provide the notice of right to opt-out" (State of California, 2020, §999.306, section c). Vermont also enacted Act "No. 171. An act relating to data brokers and consumer protection" in May 2018 (State of Vermont, 2018). However, these acts do not extend to non-commercial reuse of data.

real names or using pseudonyms, by removing any potentially identifying specifics about their lives and experiences, or amalgamating or aggregating data (A. Clark, 2006; S. L. Garfinkel, 2015; Heaton, 2004). However, some qualitative researchers describe problems that may arise from the deidentification process (N. Fielding, 2004; Hammersley, 1997; Sieber, 1991a; Stenbacka, 2001; Thorne, 1994). First, deidentification should be even more thoroughly conducted for data from vulnerable populations such as prisoners, children, people involved in illegal activities, or respondents from marginalized and minoritized communities such as Black, Indigenous, LGBTQIA+, or disabled communities. Participants from these communities may face high risk if the deidentified data are able to be reidentified (Rothstein, 2010). Smaller, more tight-knit communities may also need more careful deidentification practices to avoid potential identification of research participants (Ellard-Gray et al., 2015).

On the other hand, Parry and Mauthner suggest that "removal of key identifying characteristics of research participants may…compromise the integrity and quality of the data, or even change their meaning" (Parry & Mauthner, 2004, p. 144). Other scholars confirm that deidentification may remove important contextual information, requires time and financial resources, and may present technical challenges in the case of audiovisual data. They also suggest that the process of deidentification may not be guaranteed to prevent deductive disclosure based on other contextual information—exactly the kind of contextual information that is necessary to understand and reuse the data in the first place (Heaton, 2004; Mauthner et al., 1998; Tsai et al., 2016).

In addition to these limitations, some argue that there are instances in which deidentification may not in fact be desirable (Moore, 2012; Turnbull, 2000). Moore considers the feminist ethics of care and giving credit, showing that many studies point to "the need for, and benefits of, a careful situated and negotiated ethical practice around naming or anonymisation" (2012, p. 338).

Data curators can support deidentification practices by providing resources and services. If deidentification is not possible or desirable, data curators can also protect privacy and

confidentiality by facilitating restrictions to data access and use. Access controls are discussed further in section 2.4.2.

2.3.2.3. Intellectual property and data ownership

Intellectual property is a key consideration for qualitative data reuse (Fienberg et al., 1985; Heaton, 2004; Mauthner et al., 1998). As the United States statute states, "copyright protection subsists . . . in original works of authorship fixed in any tangible medium of expression" (U.S. Code § 102 - Subject Matter of Copyright, 1990). This means that research participants hold copyright over their own qualitative responses, and copyright holders have exclusive rights to distribute and use their works. As Mannheimer et al. write, "per this form of intellectual property protection, when someone else holds the copyright in some of a scholar's data and she was not legally assigned that right, her ability to grant others access to those data may be limited" (2019, p. 655). In order for researchers to publish the text of research participant responses, participants may need to either waive their rights or license their responses for use in the research study (Parry & Mauthner, 2004). A data use agreement or licensing agreement outlines the rights, responsibilities, and obligations of the original and secondary researchers, and may include "a description of the data that were accessed (eg, interviews, demographic data), method of access (ie, via computer software), and provisions for reference citations in publications and presentations" (Szabo & Strang, 1997, p. 72).

While such licensing could be organized as part of a research study, if no license or other permission exists, the "fair use" exemption offers a potential venue for future researchers to reuse qualitative data. According to Hirtle, Hudson, and Kenyon,

> Fair use. . . ensures that the balance between the interests of copyright owners and users can be maintained and that copyright law does not stifle the very creativity it is intended to foster. On a very practical level, it provides important protections to libraries, archives, and nonprofit educational institutions. When those organizations have a reasonable belief that their use of a copyrighted work is a fair use, many of the most stringent remedies in copyright law cannot be applied. (Hirtle et al., 2009, p. 89)

The fair use exemption is an important one for researchers reusing qualitative data, whose purpose in using the data is likely to be scholarly or educational, and for non-commercial purposes.

How researchers address intellectual property and data ownership may vary according to how and where the data were collected. For example, when collecting data from Indigenous communities, additional considerations come into play, such as the CARE principles (Carroll et al., 2021) and the First Nations principles of ownership, control, access, and possession (OCAP) (FNIGC, 2010). As Carroll et al. write, "The idea that specific communities could contribute to the development of protocols that inform the ethical use of data about them resonates with the CARE Principles, addressing concerns about fairness, trust, and accountability that are increasingly being advanced and by allowing contributors, as collectives, to have a say in how their data actually gets used" (Carroll et al., 2021, CARE in the Context of Scientific Data section, para. 7).

A 2021 report on the state of open data suggests that "copyright and licenses continue to be the area requiring the most help and have been so since the question was first asked in 2018" (Simons et al., 2021, p. 10). Data curators can advise researchers on data licensing for shared data; they can also help researchers with rights clearance, rights management, and data citation to support qualitative data reuse (Cox et al., 2017). Data curation strategies are further discussed in section 2.4.

2.4. Data curation to support qualitative data reuse

The qualitative research community and the data curation community have developed curation and archiving practices that respond to the issues described above. While these practices cannot address every issue, they do provide a set of strategies to support ethical, legal, and with epistemologically sound qualitative data reuse.

2.4.1. Metadata and documentation standards

Metadata are important for facilitating the reusability of qualitative research data. As Sieber (1991a) writes,

> Before standards of data documentation were developed, misunderstanding and unhappy outcomes were likely to mar data sharing relationships. Now, data sharing standards … can solve this problem and three others: (a) The description enables others to understand the data. (b) It allows the initial investigator to return to the data long after the needed details have faded from memory. (c) It forces the initial investigator to be systematic and rigorous in understanding the limitations of the data (e.g., details of the sampling procedure, reliability of the instruments, details of the original research design, and any deviations). And (d) it provides a basis for more systematic building on a sample, a procedure, or a body of knowledge. (p. 3)

Metadata and contextual information facilitate qualitative data reuse by those who were not originally involved in the data collection, and they serve to prevent "serious misinterpretations and biases in analysis" (White, 1991, pp. 57–58), or secondary researchers making "bolder claims than they otherwise might" (Fienberg et al., 1985, p. 7). Contextual documentation could include field notes, research diaries, correspondence, and methodological information (Corti & Thompson, 1998; Fink, 2000; Heaton, 2004). According to Corti, "for archives, documentation of the research process provides some degree of the context, and whilst it cannot compete with being there, field notes, letters and memos documenting the research can serve to help aid the original fieldwork experience" (Corti, 2000, section 6.2., para. 4). White suggests that researchers should prepare highly explicit codebooks to help future users replicate the coding process. These codebooks should contain "information on everything known about the reliability, validity, and coding problems of specific variables, extensive coding notes on problematic individual cases, page references to and quotes from the original ethnographic sources from which the coding inferences were made, plus multiple codings wherever they were done and multiple measures of the same variables wherever possible" (White, 1991, p. 54). Irwin and Winterton suggest providing seven types of information to facilitate effective data reuse: (1) citations for publications that draw on the archived data; (2) an overview of the research

design; (3) sampling decisions and how they relate to the research questions; (4) an overview of what data is provided as part of the project; (5) a descriptive profile of each participant; (6) relevant contextual information; and (7) proposed areas within the data that might warrant further analysis (Irwin & Winterton, 2011, pp. 18–19). Hinds et al. especially emphasize documentation as a mechanism for helping future researchers "feel close to a condition of 'having been there' and to imagine the emotions and cognitions experienced by the participants and the researchers during data collection and analysis" (1997, p. 414). Applying principles suggested by Chin and Lansing (2004), Faniel, Frank, & Yakel (2019) asked researchers about the different types of contextual information that they are looking for when reusing research data. In order to facilitate reuse, researchers discussed the importance of three types of contextual information: (1) data production information, including information about data collection, specimen and artifact details, the data producer information, data analysis methods, any missing data, and research objectives; (2) repository information, including provenance, reputation and history of the repository, and curation and digitization activities; and (3) data reuse information, including prior reuse, terms of use, and guidance on reuse. Initiatives such as Open Context (Kansa & Kansa, 2018), and the Data Curation Network (Johnston et al., 2018) help researchers and data repositories create documentation for qualitative research that enhances contextual integrity for data reuse.

In 2000, Corti raised several open questions regarding metadata standards for qualitative data: "Are the existing standards for study description for numerical datasets adequate? How do the emerging document type definition standards for data suit qualitative data? Do they need to be extended or reworked? At the same time, how relevant are standards adopted by the "traditional" and library communities for more complex qualitative material?" (Corti, 2000, section 7). In the years since Corti asked these questions, several initiatives have been developed to support metadata for qualitative data. The Data Documentation Initiative (DDI) (DDI Alliance, 2019) was initially created to create standardized metadata for quantitative social science data, but DDI metadata can be applied at the study level to describe qualitative research. Mannheimer et al. describe issues that may complicate the application of DDI metadata to qualitative data, including "complex study designs and relationships between files, the need to preserve the hierarchical structure of codes, and the attachment of comments or memos to specific segments of text

or to codes" (Mannheimer et al., 2019, p. 652). The Qualitative Data Exchange Schema (QuDEx), maintained by the UK Data Archive, "allows users to discover, find, retrieve and cite complex qualitative data collections in context" (UK Data Archive, 2019). QuDEx works in complement with DDI, and it incorporates object and sub-object-level metadata in addition to study-level metadata. Other context-enhancing features include: provision of highly structured and consistently marked-up data; rich descriptive metadata for files (e.g., interview characteristics, interview setting, type of object); logical links between data objects—i.e., text to related audio, images, and other research outputs; preservation of references to annotations performed on data; and incorporation of common metadata elements that enable federated catalogs across providers (UK Data Archive, 2022). In 2016, Evers called for a common exchange format to support interoperability between proprietary qualitative data analysis software (QDAS) or Computer assisted qualitative data analysis software (CAQDAS), such as NVivo and Atlas.ti (J. C. Evers, 2018); in 2019, the Rotterdam Exchange Format Initiative (REFI) released a QDA-XML format to support such interoperability. This format also has the potential to support long-term use of datasets into the future (di Gregorio, 2019). The Text Encoding Initiative (TEI) is another widely-used standard for describing textual documents (TEI Consortium, 2019). Datatags are another initiative that supports qualitative data sharing; datatags specify security and access requirements for sensitive data and attempt to reduce the complexity of data security and access by streamlining down to a few categories (i.e., "tags") (Sweeney et al., 2015).

In addition to standardized metadata, data repositories—especially social science-focused repositories such as UK Data Archive, ICPSR, and QDR—encourage researchers to include any additional materials or information that could provide context to research data. This could include documentation about research methods and practices, consent form(s), IRB approval number, information about the selection of interview subjects and interview setting, instructions given to interviewers, data collection instruments, steps taken to remove direct identifiers in the data, problems that arose during the selection and/or interview process and how they were handled, and interview roster (ICPSR, 2012). The Annotations for Transparent Inquiry initiative supports contextual information and cross-linking. Possible annotations include: excerpt from a textual source (e.g., an excerpt from the transcription for handwritten material, audiovisual material, or material generated

through interviews or focus groups); source excerpt translation; analytic note (i.e., discussions that illustrate how the data were generated and/or analyzed and how they support the empirical claim or conclusion being annotated in the text); a link to the data source; and the full citation for an excerpted source (Karcher & Weber, 2019). Qualitative Data Repository publicly released their curation handbook in 2021 (Demgenski et al., 2021), and the handbook provides guidelines for contacting and interacting with the data depositor; file processing procedures; data-level and project-level metadata; terms of use, access conditions, restrictions, and permissions; publication; and post-publication procedures.

2.4.2. Data repositories as infrastructure for sharing qualitative data

Qualitative researchers are increasingly being asked to document and archive their research data. Notably, the latest data sharing policy from NIH broadens the scope of projects that will be asked to provide data sharing plans (National Institutes of Health, 2020). The data sharing and data management plans required by funders like NSF and NIH generally include information about metadata to support future data use, as well as information about how the data will be publicly shared. These funder data sharing requirements have driven an increased demand for data curation and data repository services. Generally, data are shared in three ways: as appendices to papers and books, upon request, or more formally via a data repository (Fienberg et al., 1985). Data repositories are a growing infrastructure to support data sharing and preservation as part of the broader context of scholarly communication. Data repository staff can encourage researchers from early stages of their projects to consider how to support findable, accessible, interoperable, and reusable (FAIR) data (M. D. Wilkinson et al., 2016), and they can provide guidance on data documentation and data licensing; supporting metadata for machine-readability, search and discovery; and ensuring long-term preservation for published datasets (Mannheimer et al., 2019). Data repositories can also provide restricted access to datasets that may not be able to be made public—for example, video data that cannot be deidentified or sensitive data that should not be widely distributed. Access to datasets can be embargoed for a period of time or fully restricted. Access and use can also be restricted via data use agreements that impose certain conditions on those who would like to access and reuse the data (Leh, 2000). Corti outlines a

few questions to ask to ensure that sensitive data are appropriately safeguarded: “Are existing data preparation procedures adequate for safeguarding participants? Should qualitative and survey data from the same study be provided together? Are the access control and vetting procedures adequate?” (Corti, 2000, section 7).

There are currently more than two thousand data repositories worldwide, according to the Registry of research Data Repositories (Re3data, 2019). Some data repositories such as Dryad Digital Repository (Dryad, 2022) and ICPSR (ICPSR, 2022) provide curation support in which professional curators work with data depositors to organize data, create metadata, and otherwise support reuse. Academic libraries also provide support for research data curation (Tenopir et al., 2014, 2017; Yoon & Schultz, 2017). Notably, the Data Curation Network in the United States brings together librarians from academic libraries to support curation for institutional data repositories (Johnston et al., 2018), and the Data Curation Network has also published several data curation primers that provide curation guidance that is applicable to qualitative data, including general primers for human subjects data (Darragh et al., 2020) and qualitative data (Castillo et al., 2021), as well as more specific primers for oral history interviews (Pryse et al., 2021), and data that have been analyzed using the qualitative data analysis softwares Atlas.ti (Corral, 2020) and NVivo (Hadley, 2020). To support healthy infrastructure and long-term preservation strategies for data repositories, initiatives such as the CoreTrustSeal Trustworthy Data Repositories Requirements help repositories meet community standards for data curation (CoreTrustSeal, 2020). The TRUST Principles are designed to complement the FAIR Principles to support trustworthy practices for archived data (Lin et al., 2020).

2.5. Chapter summary

The scientific community is increasingly championing research data reuse. Qualitative data sharing and reuse has steadily grown in the late 20th and early 21st century, but several key ethical, legal, and epistemological issues arise when sharing qualitative data, including issues of context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. Data curation practices (including data curation support from data repositories and academic libraries) can help to

mitigate some of these issues, and several initiatives are in place that offer services addressing qualitative data curation and sharing. In the next chapter, I discuss issues in big social research. In chapter 4, I comparatively review the issues related to qualitative data reuse and big social data research and consider how data curation can serve as a means to help mitigate some of the epistemological, ethical, and legal issues that are present with both data types.

Chapter 3. Literature review - Big social data

In this section, I define the concepts of *big social data* and *big social research*. I then provide an overview of the history of big social research and I review the benefits of big social research. I then detail the issues that arise when conducting research with big social data, including epistemological, ethical, and legal issues, and I discuss how data curation practices can support epistemologically sound, ethical, and legal big social research.

3.1. Defining big social data and big social research

3.1.1. Big data

Big data are often defined in terms of three "Vs": volume, velocity, and variety (Diebold, 2012; Kitchin, 2014; Laney, 2001; Zikopoulos, 2012). That is, big data have large volume—comprising terabytes or petabytes of data; they have high velocity—the data are being created continually in real-time; and they exist in a variety of formats and types—big data may be structured metadata or unstructured text, audio, or video. boyd and Crawford (2012) offer additional defining characteristics for big data, writing:

> We define Big Data as a cultural, technological, and scholarly phenomenon that rests on the interplay of
>
> - Technology: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.
> - Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
> - Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy. (p. 663)

boyd and Crawford's definition helps to explain the cultural phenomenon that big data have become in our society. As big data and big data analytics have grown during the 21st Century, social scientists have begun to consider the implications of such data on social science. In 2007, Savage and Burrows suggested that, in an era where "circuits of information proliferate and are embedded in numerous kinds of information technologies" (2007, p.886), "sociologists [should] renew... their interests in methodological innovation,

and report... critically on new digitalizations" (2007, p. 896). The vast scale of big data has captured the imagination of private and public realms, leading to an era of widespread data-driven decision-making in nearly every industry, including business (e.g., Chen et al., 2012; Liebowitz, 2013; Raguseo, 2018; Schroeder, 2016), healthcare (e.g., Chawla & Davis, 2013; Raghupathi & Raghupathi, 2014; Viceconti et al., 2015; Wang et al., 2018), education (e.g., Nazarenko & Khronusova, 2017; Picciano, 2014; Williamson, 2017), and journalism (e.g., Borges-Rey, 2016; Gray et al., 2012; S. C. Lewis, 2015).

3.1.2. Big social data

The term *big social data* (or *big behavioral data*) is used to describe big data that informs social research. The definition of *big social data* specifically includes the human traces that are inherent in big data. As Amer-Yahia et al. write, "Human participation [in the creation of big data] can be *direct* such as when entering User Generated Content in blogs, microblogs, and review sites, or when knowingly participating in a crowdsourcing marketplace such as Amazon Mechanical Turk. People can also participate in *indirect* ways, simply by going about their on-line lives, when searching, reading content, shopping, or playing on-line games" (Amer-Yahia et al., 2010, p. 1259). Big social data are human-generated data, including data that result from direct human interaction as described by Amer-Yahia et al., which usually take the form of unstructured or semi-structured data such as text, videos, and audio that are created and shared online (Olshannikova et al., 2017), as well as the data that result from *indirect* human interaction as described by Amer-Yahia et al., which usually take the form of structured metadata that reflects user behavior such as interactions with interfaces, or the spatial or temporal aspects of user behavior (Gandomi & Haider, 2015). Olshannikova also categorizes big social data into three kinds—digital self-representation data, technology-mediated communication data, and digital relationships data. The chart below provides an overview of big social data, with definitions and examples; based on my reading of the literature, I include metadata as an additional kind of big social data (Drakonakis et al., 2019; Ramasamy et al., 2013; Yanai, 2012).

Table 3. Kinds of big social data that result from human interaction (adapted from Olshannikova et al., 2017)

	***Direct* human interaction data**		***Indirect* human interaction data**	
Kind	Digital self-representation data	Technology-mediated communication data	Digital relationships data	Metadata
Def-in-ition	Data related to identity depiction and communicative body in digital environment	Data related to two-way communication, knowledge creation and distribution through technology	Data that reveal digital social relationships patterns	Automatically-generated information about social posts
Ex-am-ples	*Profile data:* (i) Login data (name/username/ email address and password); (ii) Mandatory data (services and application required data, for example, full name, citizenship, birthday); (iii) Extended data (profile pictures, education, tags of interests) *Self-published content* (e.g., personal documents, pictures, videos, interests): (i) Disclosed data (to the public); (ii) Entrusted data (content sharing within trusted digital community) *Data published by the community* (e.g., pictures, narrations, videos, posts): Relates to content shared by other users, which contribute to the digital identity creation	*Private communication data:* instant 1-to-1 messaging and content sharing; *Public communication data:* 1-to-many messaging, commenting, information contribution and editing of existing entries; *Collaborative communication data:* many-to-many participatory content sharing, chats, video-conferences	*Explicit data:* Friendship data—followee/ follower data, number of likes *Implicit data:* Data, which is revealed through technology-mediated communication data (e.g., tweets could be analyzed to infer connections between people)	Timestamps, geospatial data, type of operating system, type of device, application used to post (e.g., a third-party app such as Tweetdeck or Hootsuite)

In addition to the table above, I also created Table 4, below. Table 4 is adapted from Table 1 in Chapter 2 (page 13), which describes kinds of qualitative data, and helps demonstrate the relationship and similarity between big social data and qualitative data. Contrasting Table 1 (kinds of qualitative data) with Table 4 (kinds of big social data) highlights three notable

differences between big social data and qualitative data. First, Table 4 does not include "physical objects," because big social data are by nature digital. Second, Bernard et al. categorize kinds of qualitative data as "small" and "large." Big social data can be collected at any scale, and "small" data can become "large" data if more of it is collected. Table 4 therefore does not include the "small" and "large" classifications from Table 1. Third, Bernard et al. categorize qualitative data into "public" and "private" data. As Nissenbaum suggests in her theory of contextual integrity (Nissenbaum, 2009), and as is discussed further in section 3.3.2.2, big social data exists in an ambiguous space between private and public; there are some contexts in which social media users expect privacy, and other contexts in which users consider their activities to be more public. Therefore, in Table 4, I have added a third column, "ambiguous," which includes data such as open Instagram posts from non-public figures that may be accessible publicly, but are designed for a limited, private audience.

Table 4. Kinds of big social data based on form, size, and accessibility (adapted from Bernard et al., 2017, p. 11)

	Public	Private	Ambiguous
Still Images	Webpages, online ads, Instagram posts from public figures, Flickr images, online art exhibits	Digital family photos or albums, digital patient x-rays, Instagram posts from private profiles	Open Instagram profile posts from non-public figures
Sounds	Podcast ads, digital songs, digital music albums, online news audio clips	Voice memos, voicemail messages, interview recordings, court hearing recordings	Digital oral histories
Moving Images: Video	Online video ads, online news footage, TikTok videos, digital films and TV shows	Personal iPhone videos, Snapchat video messages	Videos posted to social media by non-public figures

Texts	Online obituaries, Craigslist ads, Twitter posts using hashtags, blogs, ebooks, news websites	Emails, Notes app lists, short responses to survey questions	Comments on other people's Twitter posts, online forum posts

Social media is a common source for big social data. This dissertation uses *social media* to describe emerging digital technologies associated with Web 2.0 (D. W. Wilson et al., 2011), that allow users to post content and interact with other people. *Social media* is a broader term than *social network site*, which is defined by boyd and Ellison as a networked communication platform in which participants "(1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system" (boyd & Ellison, 2007, Social Network Sites: A Definition section). The broader term *social media* includes a wide range of digital platforms, including not only social network sites but also blogs, microblogs, photo-sharing sites, video-sharing platforms, social news and gaming, review sites, online forums, social search and crowd sourcing services, collaboration services, and virtual worlds (Ishikawa, 2015; Olshannikova et al., 2017). The uniting thread among social media platforms is that social media allows users to communicate among communities and to create and share digital content in a networked environment (Bechmann & Lomborg, 2012; Ip & Wagner, 2008; Kim et al., 2010; Lüders, 2008; D. W. Wilson et al., 2011). Bechmann & Lomborg outline three characteristics that are commonly emphasized when considering social media as a social phenomenon:

1. Social media facilitates direct communication between users—that is, communication is "de-institutionalized";
2. Users create and share their own content such as text, photos, and videos, in addition to sharing traditional published content;
3. Social media platforms are interactive and networked (Bechmann & Lomborg, 2012).

A fourth consideration is that social media platforms are often controlled by private, for-profit companies (Driscoll & Walker, 2014). Blog platforms like SquareSpace and WordPress, microblogs like Twitter, photo-sharing sites like Flickr (owned by Yahoo),

video-sharing sites like YouTube (owned by Google), online forums like Reddit (owned by Conde Nast) and Quora, virtual worlds like Second Life, or the communities that form among videogame users—these platforms all act as intermediaries between the human communities that are formed online (Fuchs, 2017; Oboler et al., 2012). All of these considerations regarding social media are therefore key considerations for researchers who collect and analyze big social data. Big social data come from an online space with specific characteristics, and access to the data is often controlled by private companies.

3.1.3. Big social research

To define big social research, I will first begin with a figure that shows two key types of internet-mediated research: *obtrusive* and *unobtrusive*, as defined by Hewson, Vogel, & Laurent (2016). Table 5 is reminiscent of Table 2 (page 15), in which Heaton (2004) gave examples of two types of qualitative data: non-naturalistic data, which are solicited for research studies, and naturalistic data, which are found or collected with minimal interference by researchers. Applying Hewson et al.'s framework, Heaton's examples of non-naturalistic data—e.g., fieldnotes, observational records, interviews, focus groups, and solicited diaries—would be characterized as resulting from obtrusive research, while Heaton's examples of naturalistic data—autobiographies, found diaries, letters, official documents, photographs, film, and social interaction—would be characterized as resulting from unobtrusive research. The table below gives examples of types of internet-mediated research.

Table 5. Types of internet-mediated research (Hewson et al., 2016, p. 37)

Type	Research strategy	Examples
Obtrusive	Surveys	● Surveys distributed via email ● Questionnaires that participants answer online
	Interviews and focus groups	● Online interviews ● Online focus groups
	Experiments	● Online experiments in which participants are aware of their participation
Unobtrusive	Observation	● Analysis of interactions in online forums and social media sites
	Document analysis	● Analysis of blogs or email archives ● Analysis of photographs on online sharing sites
	Experiments	● Online experiments in which participants are not aware of their participation

Table 5 shows the broad scope of internet-mediated research. Big social research is a sub-field of internet mediated research, and it is almost always conducted using unobtrusive methods (Bright, 2017). Additionally, while researchers can use subsets of data from online sources to conduct traditional, human-coded content analysis (e.g., Ruthven et al., 2018), conversation analysis (e.g., Paulus et al., 2016), and online ethnographies (e.g., Caliandro, 2018), big social research is by definition large-scale. Big social research is therefore commonly conducted using computational social science methods. Computational social science is a "research area at the intersection of computer science, statistics, and the social sciences, in which novel computational methods are used to answer questions about society" (Mason et al., 2014, p. 257). Computational social science began in the 2000s, and it uses methods such as natural language processing, sentiment analysis, network analysis, artificial intelligence, and deep learning techniques to draw conclusions from big social data (Bankes et al., 2002; Berkout et al., 2019; Mason et al., 2014).

3.2. History and benefits of big social research

Big social research can be traced back to social network analyses in the early part of the 20th century (Halavais, 2015; Moreno, 1934; Simmel, 1955). As archived social science data

became more common, these data were used to support larger-scale longitudinal studies (Holland et al., 2006; Neale & Bishop, 2012). However, the advent of the web and social media brought an entirely new scale to social research (González-Bailón, 2013). Big social data are now easily collected en masse by scraping the web or by using Application Programming Interfaces (APIs). Facebook and Twitter are commonly mined for social research, due to their high numbers of users and the historical ease of data collection from these platforms via public APIs. A literature review in 2012 showed exponential growth in academic research studies of Facebook during its first few years—from a single study in 2005 to 186 studies in 2011 (R. E. Wilson et al., 2012). Building on the work of boyd (2013) and Williams et al. (2013), Zimmer and Proferes (2014) demonstrate a similar growth in Twitter research—from two studies in 2007 to 382 studies in 2013. Big social research has continued to expand since then, and big social data analysis has been used to produce research across various disciplines, touching on a wide variety of topics. For example, in public health, researchers have analyzed the role of community influencers in discussions of diabetes on Twitter (Beguerisse-Díaz et al., 2017); have used sentiment analysis to understand the conversation around marijuana on Twitter (Cavazos-Rehg et al., 2015); have conducted network analysis to understand tweets about the potential contagion effect when people disclose suicidal ideation (Colombo et al., 2016); and have used content analysis of online forum posts to understand the information needs of young mothers (Ruthven et al., 2018). Notably, a literature review aiming to understand the nature of health-related research on social media found that social media is often used to reach vulnerable populations that traditionally have been more difficult for researchers to access; the study concludes that "there is a compelling need for resources designed to support ethical and responsible social media-enabled research to enable this research to be carried out safely" (Nebeker et al., 2020, p. 1). In political science, researchers have presented voting mobilization messages to Facebook users, finding that such messages "directly influenced political self-expression, information seeking and real world voting behaviour" for the targeted users, as well as other members of their social networks (Bond et al., 2012, p. 295); and machine learning and social network analysis have been used to understand political homophily on Twitter (Colleoni et al., 2014); and a traditional telephone survey investigated the extent to which social media influences political attitudes and democratic participation (W. Zhang et al., 2010). Other big social researchers have mined hashtags to investigate how Twitter is used

as a community organizing tool (Segerberg & Bennett, 2011). A systematic review of big social research in environmental science found that “this new data source offers unprecedented opportunities to extend the scope, scale and depth of research, especially insofar as the interactions between humans and the environment are concerned, but, at the same time, presents environmental researchers with a range of issues involving potential biases, big data management and rapidly evolving frameworks with which they are generally not familiar” (Ghermandi & Sinclair, 2019, p. 43). Big social data has also been used for market and brand research, investigating how social media influencers can impact brand reputations by exposing a few hundred Twitter influencers to either positive or negative tweets (Barhorst et al., 2019), and using machine learning to study the varying effects of textual and image-based brand messages across social media platforms in order to help brands develop effective strategies for social media marketing (Villarroel Ordenes et al., 2019).

3.2.1. Benefits of big social research

In a provocative 2008 editorial, Chris Anderson—then-editor-in-chief of Wired Magazine—suggested that big data would revolutionize social science methodology. “Out with every theory of human behavior, from linguistics to sociology,” he wrote. “Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves” (Anderson, 2008, para. 7). While Anderson uses heightened rhetoric to make his point, many others have acknowledged the potential of big data to reveal patterns of social behavior that could not previously be identified (Cappella, 2017; Fan & Gordon, 2014; Lazer et al., 2009; Oboler et al., 2012). Baram-Tsabari et al. write that big social research provides a “great methodological advantage: it can take what was once invisible and private and make it reachable and researchable” (2017, p. 100). Or as Bright writes, the phenomenon of big data “has quantified certain social activities that previously have been very difficult to study systematically” (Bright, 2017, p. 126). Building off of this key benefit, conducting big social research has several additional potential benefits.

Online platforms allow researchers to reach much larger numbers of participants than would be possible in traditional research, thus greatly increasing sample sizes and potentially facilitating the study of traditionally hard-to-reach populations (Taylor & Pagliari, 2018; Taylor & Moorhead et al., 2013; Baram-Tsabari et al., 2017). The large scale of big social data also allows researchers to identify and analyze trends and associations (Paul and Dredze, 2011) and supports large-scale longitudinal research over time (Taylor & Pagliari, 2018; Baram-Tsabari et al, 2017; Hokby et al., 2016). Additionally, big social data are cost-effective (Taylor & Pagliari, 2018; Munson et al., 2013). As Bright writes, "Big data are often cheap and rapid for social scientists to employ. […] This implies that theory and hypotheses can be tested more rapidly and more widely than was previously the case, in more social contexts and with fewer resources" (Bright, 2017, p. 126). Lastly, some suggest that big social research is less likely to reflect bias—such as social desirability bias—since big social research does not require direct contact between researchers and participants. For example, big social research often relies on tracking what participants say or do, rather than asking participants to respond directly to interview or survey questions (Taylor & Pagliari, 2018; McKee, 2013). According to Baram-Tsaari et al., "Mining the actual activity of users is much more reliable and accurate in revealing general social interests and needs, particularly when it comes to sensitive issues, such as online dating preferences or health-related search queries" (2017, p. 102).

All of these benefits support the increasing use of big social data to investigate human behavior. However, big social data also highlight several issues and challenges. boyd and Crawford's inclusion of "mythology" in their definition of big data (see section 3.1.1.) addresses the widespread embrace of big data as a knowledge source. In fact, boyd and Crawford respond directly to the Anderson editorial mentioned at the beginning of this section, writing, "Do numbers speak for themselves? We believe the answer is 'no'" (2012, p. 666). Kitchin elaborates on this idea, writing, "Whilst data can be interpreted free of context and domain-specific expertise, such an epistemological interpretation is likely to be anaemic or unhelpful as it lacks embedding in wider debates and knowledge" (Kitchin, 2014, p. 5).

Puschmann (2017) identifies issues that arise when researchers use data that were not originally collected for research purposes, writing, “All data need interpretation, but appropriating content created for other purposes than research is inherently risky. … Judging people by the digital traces that they leave behind is different from following a physical trail” (2017, p. 97). Most recently, the Association of Internet Researchers’ Ethical Guidelines discuss the theories that support “the propositions that digital data cannot be expected to speak for themselves, that data do not emerge from a vacuum, and that isolated data on their own should not be the end goal of a critical and reflexive research endeavour” (Franzke et al., 2020, p. 70). Section 3.3. discusses these and additional concerns in more detail.

3.3. Issues in big social research

Salganik (2018) suggests that big data have several characteristics that can be problematic for social research: they tend to be “incomplete, inaccessible, non-representative, drifting, algorithmically confounded, dirty, and sensitive” (p. 17); in other words, far from a simple solution to measuring human behavior. Puschmann (2017) emphasizes the man-made element of data, writing that data do not “simply come into being by [themselves], but [are] either the result of a planned process of elicitation or of purposeful sampling. Such processes are often made to appear more straight-forward in the ideal environment of a text book or an introductory methods class than they turn out to be in actual research” (p. 99). Proferes’ response to Puschmann further outlines the idea that data cannot “speak for themselves,” citing Barad (2003), who argues that “techno-scientific discursive practices involving language, measurement, and materiality *produce* phenomena, creating an artificial separation between researcher and the knowable” (Proferes, 2017, p. 114). Manovich also suggests that an empiricist vision of big data is misguided; he outlines several concerns in response to the rise of big social research, including data access, data authenticity, and the depth of research that is possible with this new form of data (Manovich, 2012).

As boyd and Crawford write, the advent of big data represents “a profound change at the levels of epistemology and ethics” (2012, p. 665). From the literature, I identify three key epistemological issues (section 3.3.1.): scale, context, and data quality. Clark et al. (2019)

conducted workshops with social scientists using big social data, and identified the following key ethical and legal issues (section 3.3.2.), which I also include as categories below: consent; privacy and confidentiality; and ownership and authorship. Clark et al. also highlight data sharing as a key issue—an issue that I discuss throughout this dissertation, and especially in section 3.4.

3.3.1. Epistemological issues

Epistemological challenges in qualitative secondary analysis relate to context, data quality, and data comparability.

3.3.1.1. Context

Halavais (2015) suggests that "when we collect data from [social media] platforms (just as when we collected data in traditional spaces), context matters" (p. 591). However, the context of a social media post may be absent or difficult to understand. Social media posts are by nature short pieces of text, taken from a larger context of personal and public life (Törnberg & Törnberg, 2018). This out-of-context effect is only compounded when data are amassed on a large scale. Writing about Twitter data, Bruns and Weller (2016) suggest that if the data are not captured and preserved in their entirety, context will be lost and the data will lose value. "By entirety," they write, "we mean the following dimensions: (1) the cultural artifact that is Twitter, with (1a) its look and feel and technical affordances over the course of time, and (1b) the broader societal context into which Twitter is embedded, including user numbers, demographics and usage practices, and (2) the Twitter data consisting of (2a) the complete collection of all user-generated content, including non-textual information and hyperlinks, and (2b) contextual information like collections of hashtags for important events or lists of usernames for important groups of users" (p. 185). Capturing all of these elements is difficult; in fact, boyd and Crawford suggest that context and meaning may never be accurately understood by big social researchers (boyd & Crawford, 2012). Communicating or collaborating with the original data creator has been suggested as a strategy for discerning the relevant context of research data (Pasquetto et al., 2019); however, when collecting data on such a large scale, contacting original data creators is extremely difficult, if not impossible. Some researchers have attempted to preserve context by combining social

media datasets with other data. For instance, business researchers combined social media data with customer profiles (Wittwer et al., 2017); others have used probabilistic models to identify demographic information such as geography and location, age, gender, language, occupation and class (Sloan, 2016); and researchers have collected both tweets and follow-on conversations in an effort to capture complete context (Lorentzen & Nolin, 2017). Data combining and data comparability are discussed further in section 3.3.1.3.

In addition to the challenge for researchers to understand the context of big social data, Marwick and boyd point out that a "context collapse" occurs even before researchers mine big social data. They write that when users post on social media, "multiple audiences [are flattened] into one" (Marwick & boyd, 2011, p. 122). Social media users may, in effect, post into a contextual void. Marwick and boyd suggest that that social media users, when attempting to represent the various facets of their lives and identities to a diverse community social media, "adopt a variety of tactics, such as using multiple accounts, pseudonyms, and nicknames, and creating 'fakesters' to obscure their real identities" (Marwick & boyd, 2011, p. 122). This presentation of self complicates the idea of authenticity and data quality, as discussed further below.

3.3.1.2. Data quality

Social media in particular presents complexities in terms of data quality. First, social media users may portray their identities differently online than they might in an academic study. Citing Ellison, Heino, and Gibs (2006), Manovich suggests that "peoples' posts, tweets, uploaded photographs, comments, and other types of online participation are not transparent windows into their selves; instead, they are often carefully curated and systematically managed" (Manovich, 2012, para. 26). Many scholars have also cited Goffman's idea of the presentation of self as applicable to online social behavior. (For an overview of the literature making this connection, see Hogan, 2010.) The idea of the "authentic" in big social data is additionally complicated by users' practice of creating duplicate accounts: a user may create different accounts representing different presentations of themself (Marwick & boyd, 2011). Authenticity is also complicated by the presence of bots that may be indistinguishable from "real" users, a problem that compounds

when research is conducted on a large scale. As Shah et al. write, these bots are "intended to mislead citizens and consumers… [by] generating comments on everything from political candidates' policy briefs to hotel accommodations' service quality" (2015, p. 9). A 2017 study suggested that between 9% and 15% of active Twitter accounts at that time were bots, including several subclasses of accounts such as spammers, self promoters, and accounts that post content from connected applications (Varol et al., 2017). Such accounts—representing different types of presentations of self or digital approximations of human behavior—introduce errors, bias, and distortion into studies with big social data, and may ultimately affect the overall validity of big social research.

Additionally, users of social media may not be a "complete" community, or representative of society as a whole. Some social media platforms such as Facebook and Twitter tend to be overrepresented in big social research due to ease of access (Rains & Brunner, 2015; Stoycheff et al., 2017; R. E. Wilson et al., 2012; Zimmer & Proferes, 2014), which could lead to biased research. As boyd and Crawford point out, "Twitter does not represent 'all people', and it is an error to assume 'people' and 'Twitter users' are synonymous: they are a very particular subset" (boyd & Crawford, 2012, p. 669). A 2020 survey of social media users found that Twitter users tend to have higher socioeconomic status and more advanced internet skills, suggesting that Twitter research may disproportionately leave out the views of less privileged members of society (Hargittai, 2020). Burgess and Bruns (2012) point out another potential issue with Twitter data, noting that the Twitter API delivers incomplete lists of posts with no way to know what may be missing. They write, "The total yield of even the most robust capture system (using the Streaming API and not relying only on Search) depends on a number of variables: rate limiting, the filtering and spam-limiting functions of Twitter's search algorithm, server outages and so on" (Technical, Political, and Epistemological Issues section, para. 6). Some researchers have attempted to create more representative datasets by blending big social data with smaller social datasets, as a way to "include perspectives that are both important to the data yet not necessarily present within it" (Croeser & Highfield, 2020, p. 673).

Puschmann (2017) identifies the problems with using data that were not originally collected for research purposes. He writes, "All data need interpretation, but appropriating content

created for other purposes than research is inherently risky. … Judging people by the digital traces that they leave behind is different from following a physical trail" (p. 97). With enough data, the numbers may be more easily manipulated."

3.3.1.3. Data comparability

Big social researchers may compare and combine data in order to enhance the representativeness of the data, enhance the context of the data, and achieve stronger results. Illustrative research projects include combining geotagged social media data with remote sensing imagery to enhance context (Jendryke et al., 2017), collecting data from several social media platforms to understand how technology influences political campaign communications (Bossetta, 2018), comparing traffic accident detection using Twitter data to traditional traffic accident detection methods (Z. Zhang et al., 2018), and combining traditional survey data with big social data (Stier et al., 2020). Combining big social data presents a variety of challenges. Stier et al. (2020) discuss the challenges of matching participants across datasets. Additionally, as discussed by Bossetta (2018) and Martí et al. (2019), social media platforms require varied data collection methods and offer different data sampling opportunities. Once data are collected, they may have different filetypes, different metadata fields, and different metadata standards, all of which make combining data more difficult, especially on a large scale. Data comparability and interoperability are discussed further in section 3.4.1.

3.3.2. Ethical and legal issues

Big social research may fall outside of the traditional protections and consent procedures that were outlined by the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979) and the Common Rule (1991) and are governed by ethics regulatory bodies such as institutional review boards (IRBs). IRBs in the United States have yet to come to a unified conclusion about ethical standards for big social research. As Clark et al. write, "Inadequate guidelines leave researchers and research ethics committees floundering in terms of assessing and responding to ethical issues associated with the use of digital data" (2019, p. 68). In 2011, Wilkinson and Thelwall proposed that big social data should be defined as "text," concluding that such data should

not be subject to human subjects review processes (D. Wilkinson & Thelwall, 2011). However, in the decade since, the human element of big social data has increasingly been recognized (Franzke et al., 2020; Metcalf & Crawford, 2016; Shilton & Sayles, 2016; Zimmer, 2018).

In 2013, the U.S. Department of Health and Human Services Secretary's Advisory Committee on Human Research Protections released a document outlining considerations and recommendations for human subjects regulations for internet research. The document proposes that "current human subjects regulations, originally written over thirty years ago, do not address many issues raised by the unique characteristics of Internet research" (2013, para. 1). As Buchanan writes, "While readying themselves for the next frame of internet research, researchers across the globe face significant regulatory changes, including the ways in which ethics review and approval is and should be sought and obtained" (Buchanan, 2017, p. xxxii). In 2018, the Common Rule was revised to begin to "grapple with the consequences of big data, such as informed consent for bio-banking and universal standards for privacy protection" (Metcalf, 2016, p. 31). As part of the Common Rule revision process, the U.S. Health and Human Services' Secretary's Advisory Committee on Human Research Protections issued recommendations regarding big data research that included suggestions that the Office of Human Research Protections (OHRP) provide guidance to IRBs regarding consent waiver standards for big data research, and that the OHRP suggest methods such as focus groups or community advisory boards that could help big data researchers identify the concerns of participant populations. (Secretary's Advisory Committee on Human Research Protections, 2015). These recommendations are a step toward regulating participant consent for big social research. However, they have not been codified into the new Common Rule; in practice, most big data research will still be classified as exempt from such requirements (Metcalf, 2016). Schneble et al. outline several issues regarding big social research that "may not be adequately covered by existing [ethical] guidelines" (2018, p. 1). They conclude that "if data science is to be conducted ethically, IRBs should not wait for the law to catch up, but should review such studies even if legislation does not mandate this" (p. 2).

In the European Union, the General Data Protection Regulation (GDPR) went into effect in 2018. Article 7 of this law is especially relevant to big social research, stating that "the request for consent shall be presented in a manner which is clearly distinguishable from other matters, in an intelligible and easily accessible form, using clear and plain language" and that "any part of such a declaration which constitutes an infringement of the Regulation shall not be binding" (Voigt & von dem Bussche, 2017, p. 272). While the GDPR is a step forward, the ramifications for big social research are still not fully clear (Greene et al., 2019; Vestoso, 2018).

The Association of Internet Researchers' most recent release of Internet Research Ethics, version 3.0 (Franzke et al., 2020), outlines initial considerations for each stage of research (including dissemination of research data, discussed further below), informed consent, protecting the researcher(s), and additional topics. It then suggests a general structure for ethical research online. The document also includes companion resources that explore research ethics for artificial intelligence and machine learning and corporate data, discuss feminist research ethics, and suggest an "impact model" for ethical assessment.

With these ethical guidelines in mind, I discuss in detail three key ethical issues: informed consent, privacy and confidentiality, and intellectual property and data ownership.

3.3.2.1. Informed consent

While terms of service for social media platforms and other online applications may include information or consent clauses that cover big social research, most users do not read the terms of service closely enough to support the conclusion that their use of the platform constitutes informed consent (Obar & Oeldorf-Hirsch, 2020). The GDPR's Article 7 provides regulations relating to consent, as described above; however, "it remains questionable whether the GDPR would in practice prevent the common 'click and forget' consent systems common to Internet interfaces" (Schneble et al., 2018, p. 2). And while the U.S. Health and Human Services' Secretary's Advisory Committee on Human Research Protections suggests that the use of community focus groups and advisory boards could be a way to "respect principles of autonomy and beneficence, and … ameliorate IRB concerns regarding

proposals for waiver of consent" (2015, Recommendation Three), these strategies are still largely untested.

Two high-profile cases of research with social media have brought social media research and consent procedures into the public spotlight. First, in 2012, Cornell researchers partnered with Facebook to study whether they could manipulate the content shown on Facebook users' "timelines"—the algorithmically-generated feeds that Facebook users scroll through—to provoke an emotional response (Kramer et al., 2014); ultimately prompting an Editorial Expression of Concern from the editors of Proceedings of the National Academy of Sciences (Verma, 2014), primarily regarding informed consent procedures. Second, the Cambridge Analytica scandal in 2018 also brought to light issues of consent when conducting big social research. The scandal began with a dataset collected through an app called "This is Your Digital Life," which was developed by a researcher at Cambridge University. By opting into using the app, over 300,000 Facebook users gave consent for the app to access their data and the data of their friends—a system that allowed the app to ultimately collect data from millions of Facebook users. Even though the data were deidentified and aggregated, "the fact that app users were able to consent to the use of their friends' data is very unusual, both in terms of research ethics and social media terms and conditions" (Schneble et al., 2018, p. 1). To add to the ethical complexity, no Facebook users consented for their data to be used beyond the purposes of the app, and Facebook's terms of service prohibited the sale of such data. Yet the developer of the app sold the entire dataset to Cambridge Analytica, a private political consulting firm. Cambridge Analytica then used the data to micro-target advertisements to United States voters on Facebook during the 2016 United States presidential election.

Various strategies have been employed to attempt to solve the issue of consent for big social research. For example, Hutton and Henderson used pop-up messages to evaluate participants' willingness to share certain types of data on Facebook (2013), and the Digital Footprints project provides software that provides structures to "ask participants (as normal procedure within qualitative and quantitative studies) if the researcher may retrieve and use the data in a specific research project" (Bechmann & Vahlstrup, 2015, para. 3). However, due to the sheer number of participants in a big social dataset, it is difficult, or even impossible,

to obtain individual consent, and those who consent may not be fully informed about research risks.

3.3.2.2. Privacy and confidentiality

The idea of which data are "private" and which data are "public" may blur in online contexts. Manovich (2012) cites Latour, who writes, "It is as if the inner workings of private worlds have been pried open because their inputs and outputs have become thoroughly traceable" (Latour, 2007, para. 6). While social media posts may be "publicly" available online, those who post on social media may still view their social media profile as, in a way, "private"—that is, they intend for their posts to speak specifically to their own online community. It may therefore be a breach of their privacy to collect and use such posts for research purposes.

When publicly sharing big social data, some researchers have argued that big social data are public by nature, and therefore that deidentification of such data is unnecessary. For example, in 2016, Danish researchers scraped profiles from the online dating service OkCupid and released the data without any attempt at deidentification (Kirkegaard & Bjerrekær, 2016), asserting that the data were "already public" and required no special privacy considerations or user consent (Zimmer, 2016). And in a study of diabetes using Twitter, the authors write, "We believe that the topic, analysis and results presented here serve the public interest and pose no risk to users. None of the tweets we analyse and reproduce here contain notable amounts of sensitive or private material. Indeed, the most prominent users in our data set also maintain other online profiles and produce tweets for public consumption" (Beguerisse-Díaz et al., 2017, p. 3).

However, increasingly, the consensus in the literature is that researchers must consider privacy when conducting big social research. Several theories of privacy are relevant to big social research. Palen and Dourish (2003) base their understanding of online privacy on Altman's privacy theory (1977), which suggests that "privacy regulation is neither static nor rule-based" (Palen & Dourish, 2003, p. 130). Reuter et al. (2019) also emphasize the fluid nature of privacy, pointing to Petronio's theory of communication privacy management

(2002) as a means for understanding privacy for big social data; this theory proposes that people are continually making new decisions about either disclosing or concealing private information. Nissenbaum's theory of contextual integrity (2009) has also been widely used to consider the nature of online privacy. Nissenbaum posits that, depending on the context, people have different expectations of privacy for their personal information. Reuter et al. provide the following overview: "Rejecting the traditional dichotomy of public versus private information, as well as the notion that a user's preferences and decisions of privacy are independent of context, [the theory of] contextual integrity provides a framework for evaluating the flow of personal information between different agents; it also provides a framework for explaining why certain patterns of information flow might be acceptable in one context but viewed as problematic in another" (Reuter et al., 2019, p. 2). As Marwick & boyd write, citing Nippert-Eng (2010), "Anthropologists and sociologists maintain that privacy is a social construct that reflects the values and norms of individuals within cultures, meaning that the ways in which people conceptualize, locate, and practice privacy varies tremendously" (Marwick & boyd, 2014, pp. 3–4). Palen and Dourish elaborate further, writing, "Privacy management is not about setting rules and enforcing them; rather, it is the continual management of boundaries between different spheres of action and degrees of disclosure within those spheres. Boundaries move dynamically as the context changes" (2003, p. 131). Ito (2008) introduces the idea of networked publics—that is, "a linked set of social, cultural, and technological developments that have accompanied the growing engagement with digitally networked media" (Ito, 2008, p. 2, as quoted in boyd 2010). Marwick and boyd (2014) extend the idea of networked publics into the concept of networked privacy. Marwick and boyd interviewed teenagers about privacy on social media and found that "to manage an environment where information is easily reproduced and broadcast, … many teenagers conceptualize privacy as an ability to control their situation, including their environment, how they are perceived, and the information that they share." Marwick and boyd propose that "just as people seek out privacy in public spaces, ... they take steps to achieve privacy in networked publics, even when simply participating in such environments requires sharing" (p. 4), ultimately defining networked privacy as the "ongoing negotiation of contexts in a networked ecosystem in which contexts regularly blur and collapse" (Marwick & boyd, 2014, p. 13). Together, these various theories of privacy suggest that people's expectations of privacy and their strategies for protecting their privacy online

are constantly changing and adapting, depending on a variety of factors, including "physical environment, audience, social status, task or objective, motivation and intention, and … [the] information technologies in use" (Palen & Dourish, 2003, p. 131).

Several studies have attempted to understand users' expectations for privacy online. A 2014 Pew Research Center study found that most respondents wanted to do more to protect their privacy, yet they also believed "it is not possible to be anonymous online" (Madden, 2014, p. 5). Reuter et al. find that "most users do not think monitoring Twitter for the purpose of clinical trial recruitment constitutes inappropriate surveillance or a violation of privacy" (p. 12). However, they also note that "the expressed attitudes were highly contextual, depending on factors such as the type of disease or health topic and the entity or person who monitored users on Twitter" (Reuter et al., 2019, p. 12). Golder et al. also concluded that participant responses to social media research varied, depending on "the type of social media platform ... the vulnerability of the social media use" (Golder et al., 2019, p. 1). Fiesler and Proferes (2018) find that Twitter users have concerns about privacy that track the themes of the Belmont Report (1979): respect for persons, beneficence (minimizing harm), and justice. Social media platforms have responded to user privacy concerns with more granular privacy-management controls (Fiesler et al., 2017; Twitter, 2022). However, the privacy settings of social media platforms generally default to open; users must opt into granular privacy controls, and users may have difficulty implementing these controls (Sleeper et al., 2013).

As in offline research, issues of privacy and confidentiality are especially important "for research involving vulnerable populations who may have limited understanding of the implications of disclosing personal information on these platforms" (K. Clark et al., 2019, p. 61). This is even more true because big social data is used by government entities and advertisers for surveillance. In 1991, Sieber wrote that surveillance "is not a legitimate use of shared data and may be damaging to science" (1991b, p. 148). However, the social media business model is to provide "free" services to users; the revenue comes from advertising dollars. This model gave rise to the "internet-age dictum that if the product is free, you are the product" (Lanchester, 2017, para. 14). As Oboler, Welsh, and Cruz write, the ad-driven business model "places the individual's interest in privacy at war with the advertisers'

interest in greater customer profiling" (Oboler et al., 2012, The Business Customers of Social Media section, para. 1). The Documenting the Now (DocNow) project has also released a white paper discussing the risk that big social data archiving could be used to facilitate or enhance police surveillance (Jules et al., 2018). DocNow is discussed further in section 3.4.2.

3.3.2.3. Intellectual property and data ownership

Big social research also raises issues about intellectual property and data ownership. In 2018, Facebook CEO Mark Zuckerberg testified before Congress, saying, "Every piece of content that you share on Facebook, you own, and you have complete control over who sees it and … how you share it, and you can remove it at any time" (Washington Post, 2018). However, in the United States, intellectual property on social media is still a relatively gray area of law (Blank, 2018; Bosher & Yeşiloğlu, 2019; Doft, 2015; Wilkof, 2016).

As noted in section 3.1.2, a key consideration for big social data is that they are often controlled by private, for-profit companies. Even if the text, image, and video content of social media posts are the intellectual property of the users who posted them, these posts are licensed to social media companies through the companies' terms of service. Such terms of service govern the behavior of users, developers, researchers, and archivists (Puschmann & Burgess, 2014), and they are a reflection of how much value and revenue are generated through user data. Companies use these data both internally for user studies and analytics, and they can additionally profit by selling their user data to data brokers and advertisers.

Because social media platforms view their data as a corporate asset, they will take steps to protect that data, much as they would any other corporate asset, by trying to limit the ability of outside entities to harvest and reuse the data. In recent years, social media companies have invoked the Computer Fraud and Abuse Act (CFAA) (the primary federal anti-hacking law) to try to prevent other companies from using automated bots to scrape data from their platforms. Some legal scholars have voiced concern that if the courts interpret the CFAA to prevent web scraping of public data, large social media companies could effectively bankrupt smaller analytics companies and research organizations through expensive legal proceedings and data

access fees, resulting in data monopolies (McRory, 2021).

A notable example of this strategy is described in the court case of hiQ Labs v. LinkedIn Corporation, (*hiQ Labs, Inc v. LinkedIn Corporation*, 2019). In that court case, LinkedIn (a professional networking platform) claimed that the CFAA prohibited hiQ (a data analytics company) from scraping the information that LinkedIn users shared on their public profiles—data that could be viewed by anyone with a web browser. The federal court tentatively concluded that the aim of the CFAA was to punish unauthorized intrusion into a computer or a computer system, but not to punish unauthorized use of information that was freely available without hacking into a system. This interpretation of the law was ratified two years later by the United States Supreme Court in a case called Van Buren v. United States (*Van Buren v. United States*, 2021). In Van Buren, the Supreme Court ruled that the CFAA does not prohibit a person from using data for unauthorized purposes, as long as the person had the authority to access that data (i.e., the authority to access the computer system as a whole, as well as the authority to access the files, folders, or databases where the data was stored). However, the Van Buren decision did not definitively resolve the question of whether web scraping is prohibited by the CFAA—because, in footnote 8 of the Supreme Court's opinion, the court declared that it was not deciding whether a third person's right of access to a social media platform's data turns only on technological (or "code-based") limitations on access, or whether instead a third person's right of access might be controlled by "[the] limits contained in contracts or policies" (*Van Buren v. United States*, 2021).

Social media terms of service may limit how much big social data can be legally re-shared by primary researchers. For example, while Twitter's API provides access to varying levels of user data, Twitter's developer terms of service stipulate that only Tweet IDs, not full-text tweets, should be published by Twitter data researchers: "If you provide Twitter Content to third parties, including downloadable datasets of Twitter Content or an API that returns Twitter Content, you will only distribute or allow download of Tweet IDs, Direct Message IDs, and/or User IDs." (Twitter, 2020)

Archives have responded by publishing "dehydrated data" (Hemphill et al., 2018)—that is, a list of Tweet IDs that represent a full Twitter dataset. These data can then be "hydrated" to

include the full text. However, because all tweets that have been deleted or protected by the user since the time the research was conducted will not surface in the “hydrating” process, such lists may have reduced value in terms of supporting reproducibility.

In the aftermath of the Cambridge Analytica scandal, many social media companies updated their terms of service and their API access in order to restrict use of data even further (Bruns, 2019). For instance, the Twitter Terms of Service for Developers suggests that:

> Prohibited uses of our data and developer products include investigating or tracking Twitter users or their content, as well as tracking, alerting, or monitoring sensitive events (such as protests, rallies, or community organizing meetings). Other categories of activities prohibited under these terms include (but are not limited to):
>
> - Investigating or tracking sensitive groups and organizations, such as unions or activist groups
> - Background checks or any form of extreme vetting
> - Credit or insurance risk analyses
> - Individual profiling or psychographic segmentation
> - Facial recognition
>
> These policies apply to all users of our APIs. Any misuse of the Twitter APIs for these purposes will be subject to enforcement action, which can include suspension and termination of access. (Twitter Developers, 2020)

By restricting the use of big social data in these ways, Twitter and other social media companies attempt to protect themselves and their users. However, in effect, these restrictions may limit the topics of study for academic researchers.

3.4. Data curation to support big social data use and reuse

Data librarians, curators, and repositories play a role in supporting curation for big social data, especially by supporting data documentation and archiving to encourage discovery, protection, documentation, and preservation of big social data.

3.4.1. Metadata and documentation

Descriptive and technical metadata are vital to the reuse of big social data. Social media contains embedded metadata. Using Twitter as an example, each tweet includes not only the plain text written by the Twitter user but also "150 pieces of metadata, such as a unique numerical ID, a timestamp, a location stamp, IDs for any replies, favorites and retweets that the tweet gets, the language, the date the account was created, the URL of the author if a Web site is referenced, the number of followers, and numerous other technical specifications" (Zimmer, 2015, Challenges for Practice section, para. 2). A second order of metadata can additionally be identified within the text of the tweet: hashtags, @-mentions, and URLs. As Driscoll and Walker write, "Taken together, these primitive components provide a set of basic descriptive characteristics that might be reported about any collection of tweets" (2014, p. 1747). However, capturing the full extent of these descriptive characteristics is difficult. Social media posts represent ongoing conversations with other users, and they contain references to live webpages and constantly updating hashtag usage. In order to fully capture the context of big social data, one must archive both the text of the post, the embedded metadata, and each of the linked resources; some archives, such as the United Kingdom National Archives' social media archive, link archived social media posts with the archived webpages that they link to; as Thomson and Beagrie write, "Preserving social media means capturing enough content to provide meaning but also finding practical solutions to managing such large, diverse, and interlinked material" (Thomson, 2016, p. 24).

Additionally, the metadata embedded in big social data vary by social media platform. As Acker and Kriesberg note, this "lack of descriptive standards will continue to impede cross-comparison of social media data without significant data wrangling and standardization efforts—there are no data models for cross-walking or mapping like-with-like across platforms, for example a tweet, a Facebook post and a YouTube video that all link to the same content or event such as a townhall livefeed" (2017, p. 7). While the proprietary nature of many social media platforms may continue to impede the development of standardized metadata that would facilitate cross-platform analysis, data sharing, and reuse, there are some models for unified metadata schemas (e.g., DDI Alliance,

2019; Schema.org, 2020) that could either be adapted or inform similar community efforts specific to big social data.

Researchers and data curators can also work together to ensure that "the objectives, methodologies, and data handling practices of the project are transparent and easily accessible" (Rivers & Lewis, 2014, Proposed Guidelines for the Ethical Use of Twitter Data section). As I write with coauthor Elizabeth Hull, "When researchers are transparent about their process, they support a culture of openness, facilitate data reuse, and help educate other researchers about methods for ethical data sharing" (Mannheimer & Hull, 2018, p. 201). Kinder-Kurlanda et al. (2017) also point out that the associated code should be archived alongside the data, and suggest that metadata standards that have been developed for social science data, such as the Data Documentation Initiative (DDI Alliance, 2019), can be adapted to document big social data as well. However, there is currently no existing metadata standard that is specific to big social data.

3.4.2. Data repositories as infrastructure for sharing big social data

Manovich (2012) outlines the idea of access as a key issue of big social data use. He writes, "Only social media companies have access to really large social data—especially transactional data. An anthropologist working for Facebook or a sociologist working for Google will have access to data that the rest of the scholarly community will not" (Manovich, 2012, para. 21). Driscoll and Walker put a finer point on the issue, writing "The stewardship of [an] unprecedented record of public discourse depends on an infrastructure that is both privately owned and operationally opaque" (2014, p. 1746). This discrepancy of access could lead to a new type of digital divide—a "big data divide" (Andrejevic, 2014), that is, a divide between those who create big data, and those who can put it to use. boyd and Crawford (2012) call these two groups "the big data rich and the big data poor;" Bruns (2013) calls them "data haves" and "data have-nots." The issue is ultimately whether social scientists can gain access to the data that they need to find insights into human behavior. Data archiving in repositories is one strategy to guarantee that researchers will have access to big social data.

As discussed in section 3.3.2.1., the 2018 Cambridge Analytica controversy highlighted the breadth of ethical questions that arise when conducting big social research, and it brought widespread public attention to the real-world consequences that can result from social media research and social media user manipulation. Perhaps most notably for academic researchers, the Cambridge Analytica scandal brought an end a "Wild West of social media research" (Puschmann, 2019), characterized by easy access to big social data, with few rules or regulations. The change was so swift and disruptive to the status quo that Bruns called it the "APIcolypse" (2019). As noted in section 3.3.2.3., social media companies refined their terms of service regarding data use and limited API access. Facebook partnered with researchers at Harvard and Stanford Universities to form Social Science One, which calls its model "a critical step toward independent analyses of the dynamics of social media's effect on society" (King & Persily, 2020, p. 5) and provides structures for academic researchers to gain extended access to Facebook data. However, these public-private partnerships still place power in the hands of the social media companies. Public data archiving could support open access to big social data for future scholarship.

Weller and Kinder-Kurlanda suggest that archives and data repositories should "fuel the discussions on: suitable documentation practices and metadata standards, different models for data access (e.g., embargoes, access to sensitive data), [and] practices for anonymization of social media datasets" (2016, p. 170). In 2010, the Library of Congress began one of the first major projects aimed at archiving big social data, partnering with Twitter with the goal of archiving all Twitter content. However, the effort was fraught with challenges related to the size, complexity, and continuous growth of the data, as well as access and query processing; access restrictions; content restrictions; privacy; and user control—with the result that the Library of Congress never provided researcher access to the Twitter content (Zimmer, 2015). In December 2017, the Library of Congress announced that they would begin to "acquire tweets on a selective basis—similar to our collections of web sites" (Osterberg, 2017, para. 4). The Internet Archive collects some social media sites and profiles, but the crawls are not comprehensive, and the crawled website snapshots are generally accessible only through search and browse—a far less user-friendly access model than the API access that social media sites provide (Ben-David & Huurdeman, 2014; Vlassenroot et al., 2019). This leaves the landscape of social media archiving as an undertaking conducted

largely on a project-by-project basis. Libraries, archives, and data repositories collect big social data according to their own collecting aims and their views of what constitute relevant topics, while individual researchers share big social datasets only in support of their published articles.

Several projects specifically address the work and challenges of harvesting and archiving big social data. A few examples are George Washington University's Social Feed Manager, ICPSR's Social Media Archive, GESIS, the US National Archives, the UK Data Service, and the Documenting the Now Project. The Documenting the Now project "develops tools and builds community practices that support the ethical collection, use, and preservation of social media content" (DocNow, 2020), and has created tools such as the DocNow Twitter appraisal tool, a "rehydrator" that pulls full tweet text from Tweet ID numbers, and a catalog that links to social media datasets in data repositories such as Dryad, Zenodo, and Dataverse. The team has also produced a white paper examining the ethics of archiving big social data (Jules et al., 2018), and created a labeling system called Social Humans (Dolin-Mescal, 2018), inspired by the Local Contexts project's Traditional Knowledge labels and licenses, which are applied by indigenous communities to communicate data ownership and access considerations for Indigenous materials. Social Humans labels aim to empower users and librarians to support ethical reuse of big social data.

Data curators and repositories traditionally seek to protect participant privacy through deidentification and access controls. However, even if researchers actively try to deidentify shared big social data, the practice of deidentification may be difficult. The 2008 Taste Ties and Time dataset was an early example of the difficulty of deidentifying big social data. In the associated study, researchers at Harvard mined the Facebook profiles of college students to investigate how their interests and friendships changed over time (K. Lewis et al., 2008). These student Facebook users were unaware that their data were being collected and used by academic researchers. The authors then openly released the "deidentified" Facebook dataset in an effort to support future research with the data; however, the data were quickly revealed to be highly re-identifiable (Zimmer, 2010). Schneble et al. (2018) further emphasize that aggregating data has the power to transform seemingly benign or "public" data into more sensitive or private data. They note that "in some situations, combinations of

public data might also lead to data being revealed that participants or identifiable groups (especially if they are vulnerable) would want to be kept private" (p. 1). They also note that "data that are anonymized today might be made re-identifiable tomorrow [through enhanced data technologies]" (p. 2). Metcalfe points to unknown risks as a result of algorithmic analysis: "The power and peril of big data research is that large datasets can theoretically be correlated with other large datasets in novel contexts to produce unforeseeable insights. Algorithms might find unexpected correlations and generate predictions as a possible source of poorly understood harms" (2016, p. 33). Speaking further about the reidentifiability of big social data, Chu et al. (2021) compare the identifiability of traditional qualitative research with that of big social research. They point out that in qualitative research studies—which must comply with traditional human subjects protections—it is common to directly quote respondents in order to support key findings and highlight ideas of interest, and it is possible for such quotes to be kept anonymous. "In contrast," Chu et al. write, "Twitter is accessible by anyone with an Internet connection; a Twitter account is not necessary to view publicly available tweets. Therefore, researchers studying social media network data must be cognizant of the degree to which their 'participants' may be discoverable" (Chu et al., 2021, p. 42). Markham suggests that this problem may be solved by "ethical fabrication" in which big social researchers rephrase social media posts to reflect the intention of the statement without quoting posts verbatim (Markham, 2012).

In order to support the privacy of the social media users, it may also be beneficial for data repositories to restrict access in the same way as they might for sensitive qualitative data, with access provided only to researchers who have been carefully vetted. Data repositories could look to existing projects such as the content management system Mukurtu (Mukurtu, 2020), which was designed specifically to accommodate the different levels of access permissions for digital objects that may be required by Indigenous communities. The ideas behind Mukurtu could act as a guide for future big social data archiving projects that require granular access permissions. Another emerging privacy protection strategy is to create data enclaves that allow users to access the data from their own computer but do not allow users to download the data or remove it from the remote server (Mathur et al., 2017). Data enclaves are also being adapted to allow researchers to conduct analysis and receive outputs

without viewing full datasets (Hemphill et al., 2018). This strategy is used for qualitative studies in which the risk of disclosure is too high even for restricted access, and the strategy is being increasingly used for big data as well (The Economist, 2022).

As big social data archiving expands, so do the challenges and uncertainties related to big social data curation. Libraries, archives, and data repositories are still in the process of developing best practices that can support legal and ethical preservation of, and access to, big social data.

3.5. Chapter summary

The advent of big social data has the potential to reveal large-scale insights about human behavior. However, several key ethical and epistemological issues arise when conducting research with big social data, as well as when sharing or archiving those data. These issues include context, data quality and trustworthiness, data comparability; informed consent, privacy and confidentiality, and intellectual property and data ownership. Data curation practices, including data curation support from data repositories and academic libraries, can help to resolve some of these issues. However, there is still little consensus about how to "manage the balance between transparency and protecting research subjects" (Sujon, 2017, p. 92).

This chapter shows that big social research has similarities to qualitative data reuse, including unaddressed issues that resemble those raised by qualitative data reuse, but these issues often have different dimensions from the ones associated with qualitative data reuse. In chapter 4, I review and compare the issues related to qualitative data reuse and big social research, with an eye toward using data curation practices as a means to resolve or mitigate some of the epistemological and ethical issues that are presented by both data types.

Chapter 4. Synthesis of issues and data curation strategies

The literature reviewed in Chapters 2 and 3 reveals that issues in qualitative data reuse and big social research are similar, but their respective communities of practice are under-connected. Both types of data present the epistemological issues of context, data quality, data comparability, and scale, as well as the ethical and legal issues of informed consent, privacy and confidentiality, and intellectual property. However, despite these similarities, big social research has not yet been widely framed as a form of qualitative data reuse, and qualitative data reuse has only begun to be discussed through a big social data lens. The literature suggests that those who reuse qualitative data and those who conduct research with big social data can benefit from the strengths of each, and that data curation strategies can be adapted to address key issues presented by both types of research.

Qualitative data reuse is a more established practice, and thus there are more developed data curation strategies to support, epistemologically sound, ethical, and legal sharing and reuse of qualitative data. Even so, issues still exist pertaining to the reuse of qualitative data. In comparison, data curation for big social data is less well-developed, and there is little consensus about how to "manage the balance between transparency and protecting research subjects" (Sujon, 2017, p. 92). This chapter synthesizes the key issues relating to qualitative data reuse and big social research, and it reviews data curation practices that support epistemologically-sound, ethical, and legal data use for these two types of data. Ultimately, this chapter poses questions that will be answered through a series of interviews with researchers.

4.1. Epistemological issues

4.1.1. Context

One key issue for both qualitative data reuse and big social data is preservation of the data's context. For both types of research, there is concern that researchers may misconstrue or fail to understand the significance of the data outside of the data's original context.

For qualitative data, these concerns center around whether the data can be meaningfully used without the knowledge and expertise of the researchers who conducted the original research project. For big social data, the problem is that individual posts are removed from their context by the very nature of the research process itself, which isolates text, photos, or other content from the larger context of the user's personal and public life. This out-of-context effect is only compounded when data are amassed on a large scale. For big social data, the researcher may never speak to the people who wrote the posts, or know their identities or broader contexts.

Marwick and boyd also refer to a "context collapse" in big social data, in which "multiple audiences [are flattened] into one" when posting on social media (Marwick & boyd, 2011, p. 122), making the context and viewpoint of big social data difficult to discern: to whom is a user speaking when they post on social media? This context collapse can also apply to archived qualitative data: while the original audience and context of the data are generally more easily identifiable, the future audience of archived qualitative data is unknown.

For both big social data and qualitative data, the literature suggests that the full context and meaning of the data may never be accurately understood by qualitative data reusers or big social researchers.

4.1.1.1. Data curation to enhance context

Communicating context for data is a key issue for both types of data. For qualitative data, data curators can encourage researchers to document contextual information throughout the research process. For example, research participants may provide information about the broader context in the course of their narratives, and researchers can be asked to provide additional contextual information when archiving the data. Collectors of qualitative research data often know more contextual information about the data they collect, because qualitative research often involves embedding in communities and working in collaboration with research participants.

For big social data, as noted above, it may be impossible for the researcher to discern the full context of the data. Contextual clues can sometimes be found in the form of embedded geolocation metadata, @-mentions, or hashtags. However, the researcher likely does not have any additional knowledge of context, other than information about their own data sampling and collection methods. Because big social data are collected on a large scale, the data may come from a wide variety of contexts, none of which may be discernible by the researcher.

4.1.2. Data quality and trustworthiness

Issues of data quality and trustworthiness take on different dimensions when considering qualitative data and big social data. For qualitative data, quality issues arise primarily from human error. Humans throughout the process can introduce errors through simple mistakes and inaccuracies. And errors can be introduced at many stages in the research—from research subjects, reporters or recorders of field data, researchers, and data coders.

Data quality issues for big social data have additional complexities that can introduce different types of errors. Because this type of research relies on automated data collection and analysis, there are fewer opportunities for simple human mistakes. However, quality issues can arise from the element of self-performance that is often present in big social data; social media users are not speaking directly to the researcher, but rather to a perceived online community. Other quality issues can result from the specific environment of online social platforms. Fake accounts and bots can introduce errors, bias, and distortion. Additionally, big social data sampling is often biased because social media APIs may not return complete data, and because users of social media platforms may not be representative of society as a whole. These sampling issues can sometimes be ameliorated by combining datasets to attempt to create a more representative set of users (see Data comparability, below).

For both types of data, systematic errors can be introduced as a result of bias. When researchers reuse qualitative data or combine datasets, these bias errors can be compounded. Nevertheless, data quality and trustworthiness is an issue that affects all data

reuse; it is not unique to qualitative data or big social data. Thus, any insights gained by comparing the research practices of these two communities—the researchers who share and reuse qualitative data and the researchers who use big social data—will only provide a starting point for addressing the problem of data bias in its larger context.

4.1.2.1. Data curation to communicate data quality and trustworthiness

Data curators can support documentation of the research process when sharing data, including identification of potential errors, potential bias, and potentially missing data. Data curators can also make sure that descriptive metadata are of high quality.

4.1.3. Data comparability

For both archived qualitative data and big social data, researchers can assess the comparability of the data by (1) identifying the extent of missing data; (2) identifying the convergence of primary and secondary research questions; and (3) assessing the methods used to produce the primary data.

The comparability of big social data is additionally affected by the issue of metadata interoperability. While standardized metadata such as Data Documentation Initiative (DDI) metadata are commonly used for qualitative data, metadata for big social datasets are less standardized. Social media platforms use different metadata schemas, and it can be difficult and time-consuming to combine multiple big social datasets if the metadata are not interoperable.

Lack of comparability is an especially important issue for both qualitative data reuse and big social research. For both types of data, combining multiple datasets would help support larger-scale studies, which is a particular focus for qualitative data, but can apply to both. Combining data could also be used as a strategy to better understand context and to enhance data quality, which is a particular focus for big social data, but can apply to both.

4.1.3.1. Data curation to enhance comparability

For qualitative data, curators can support comparability by encouraging researchers who publish qualitative data to include clear documentation addressing missing data and describing their research questions and methods. For big social data, curators can adapt existing standards such as DDI and Qualitative Data Exchange Schema (QuDEx) to support better data comparability—by adapting these standards to better fit big social data, and by combining them with other standardized metadata schemas that are used on the web, such as W3's Schema.org metadata. Additionally, the research and data curation communities could advocate for interoperable metadata standards that can be adopted by the social media platforms themselves.

4.2. Ethical and legal issues

The ethical and legal issues identified by the literature review are only specifically regulated by IRBs for primary qualitative data collection. Most archived qualitative data and big social data are categorized as publicly available, and therefore exempt from IRB oversight. Moreover, IRB regulation was originally developed for biomedical research, and is less well-attuned to social science human subjects research. Because the 2018 revision of the Common Rule declined to regulate big social data, researchers are left to evaluate the ethics of their practices for themselves, without official regulatory guidance (Cooky et al., 2018).

Resources such as AoIR's Ethical Guidelines (Franzke et al., 2020), professional working groups such as Force11/COPE Research Data Publishing Ethics working group (Puebla & Lowenberg, 2021), and organizations such as the International Data Spaces Association (IDS Association, 2022) all point toward an emerging infrastructure to support ethical and legal data practices in qualitative data reuse and the big social research. However, this dissertation suggests that our communities of researchers and data curators need to continue to develop specific standardized practices for ethically using, sharing, and reusing qualitative and big social data.

4.2.1. Informed consent

The issue of informed consent applies similarly to qualitative data reuse and big social research. While researchers increasingly include language in consent agreements regarding data reuse, it is impossible for research participants to anticipate the full scope of potential reuse of open data. Ethical questions will therefore inevitably arise regarding whether truly informed consent is possible for either qualitative data reuse or big social research.

Social media terms of service often include user agreements with language about the use of the data for research purposes. However, users generally do not read terms of service closely, and even if they do, the extent of future data reuse is impossible for them to determine or foresee. A similar problem of consent arises when qualitative data is reused. However, the participants who provided the data for a qualitative dataset at least spoke with the researchers and consented to the original study, whereas the "participants" in big social data studies may not even be aware that they are participants.

The U.S. Health and Human Services' Secretary's Advisory Committee on Human Research Protections (SACHRP) recommends convening focus groups or community advisory boards in an effort to "ameliorate IRB concerns regarding proposals for waiver of consent" (2015, Recommendation Three). As Metcalf writes, "the high standard of informed consent is intended primarily for medical research, and can be an unreasonable burden in the social sciences. However, to default to end user license agreements poses too low a bar... [E]xplicit guidelines and processes for future inquiry and revised regulations are warranted" (2016, p. 33).

Obtaining informed consent is challenging both in qualitative data reuse and big social research. In the case of deidentified qualitative data that have been shared for the purpose of reuse, participants often cannot be contacted to obtain their informed consent for new research. And in the case of big social data, the scale of the data makes it close to impossible to obtain informed consent from each participant.

Discussions of consent in both qualitative data reuse and big social research often emphasize the value of big social data and data reuse, which leads ethics regulatory bodies and researchers to try to find strategies that support new forms of consent as alternatives to the traditional, limited definition of informed consent. The 2018 revision of the Common Rule codifies the idea of broad consent, and the SACHRP suggests additional strategies for determining whether certain user groups would be likely to consent to big social research, without the need to contact individual users from big social datasets. However, the question of informed consent, especially for qualitative data (including big social data), continues to be a thorny one. In particular, when research involves sensitive topics or vulnerable populations, the format and content of the participants' consent must be given careful consideration, and the data should be scrutinized for potential identifiability. (See next section for further discussion of identifiability.)

4.2.1.1. Data curation

Data curators most often come into the research process after the data have been collected. However, if contact between researchers and data curators is initiated early in a research study, data curators can help draft broad consent language that supports foreseeable data reuse; they can also encourage convening focus groups and advisory groups to provide community guidance on consent; and they can suggest the use of automated strategies for obtaining consent from users when researchers collect big social data.

4.2.2. Privacy and confidentiality

While privacy and confidentiality are major issues for both qualitative data and big social data, these two types of data present distinct concerns regarding privacy and confidentiality.

One problem in qualitative data reuse is that measures designed to achieve deidentification may compromise the integrity and quality of the data, or may remove important contextual information. The flip side of this problem is that even careful deidentification is not guaranteed to prevent deductive disclosure of participants' identity based on the contextual information that is provided to support data reuse.

For big social data, deidentification is difficult, if not impossible (see Zimmer, 2010). Some social media platforms are full-text searchable, which means that any exact quote could disclose a user's identity; in addition, the large scale of big social data makes it easier to deduce identities, therefore putting participants at risk.

A unique consideration for big social data is that, while social media posts may be "publicly" available online, users may still view their social media posts as private because they intend to speak specifically to a personal online community. It may therefore be a breach of privacy to read, collect, and use such posts for research purposes.

4.2.2.1. Data curation

Data curators can provide guidance and/or employ deidentification procedures during the curation process. These procedures include deleting names or replacing them with pseudonyms, removing potentially identifying details about participants' lives and experiences, and amalgamating or aggregating data. When data cannot safely be deidentified (or safely shared without deidentification), repositories can impose restricted access—either by embargoing data for a period of time or by providing access controls for the data. Data use agreements dictate the conditions required for other researchers to access and reuse the data.

4.2.3. Intellectual property

As discussed in Chapter 2, qualitative data are the intellectual property of the research participants. Thus, in order for researchers to publish the text of participant responses, participants need to either waive their rights or license their responses for use in the research study. In addition to authorizing the use of the participants' responses in the primary study, licensing agreements can also outline the rights, responsibilities, and obligations of future researchers using the data. The doctrine of fair use may apply to qualitative data, since reuse is generally for scholarly or educational non-commercial purposes. However, from a data curation perspective, the clearest strategy to address intellectual property concerns is to apply a license that supports reuse of the data. For this reason, there is growing support in the data curation community to release data into the

public domain using the CC0 public domain waiver from Creative Commons (Creative Commons, 2014; Schaeffer, 2011; Schofield et al., 2009). If data are released into the public domain, continuing reuse becomes simpler and rights management more straightforward.

Big social data sharing is made more complex by the fact that these data are often controlled by private for-profit companies. Even if the contents of social media posts are the intellectual property of the users who posted them, social media companies may still implement terms of service that govern the behavior of users, developers, researchers, and archivists. This may prevent sharing big social data in the ways that qualitative research data would be shared. One example of data sharing restrictions is the case of Twitter, whose Terms of Service dictate that only Tweet ID numbers may be openly shared (for further information, see section 3.3.2.3.). In response, tools have been developed, such as DocNow's Hydrator tool, which uses the Twitter API to pull complete metadata for shared Tweet IDs (Summers, 2017).

4.2.3.1. Data curation

For qualitative data, if data curators can reach researchers early in the process, they can encourage researchers to include data licensing considerations in their initial consent agreements. Data curators can also help researchers choose a data license at the data archiving and sharing stage. Once data are shared, data curators can help future users of the data understand intellectual property rights management—how users can and cannot reuse shared data.

For big social data, data curators can help researchers navigate the social media platforms' terms of service so that they can collect, archive, and share data in accordance with these terms. Data curators can also encourage researchers to include tools such as the Twitter Hydrator as part of the data deposit, to support usability for the archived data.

4.3. Summary of similarities and differences

The issues described above are all key to successful qualitative data reuse and big social research. However, some issues may be thornier than others—that is, some issues have

larger potential consequences, some issues may include more potential to harm participants, and some issues may be more difficult to resolve or alleviate. For both qualitative data reuse and big social research, epistemological issues may lead to less accurate or reduced-scale research, and negative consequences could include harm to researchers' reputations within the scholarly community, or reduction in the overall usefulness of their research results. On the other hand, ethical and legal issues can result in consequences that extend beyond the scholarly community, including litigation against institutions, harms to participants, and negative publicity (e.g., Mello & Wolf, 2010; Verma, 2014). Table 6 provides a broad overview of the similarities and differences between these issues in the contexts of qualitative data reuse and big social research, and Table 7 describes data curation strategies for each issue.

Table 6. Similarities and differences of issues in qualitative data reuse and big social research

Issue	Similarities	Differences
Context	● Data may not be properly understood outside of their original context. ● Original researchers can provide context information via metadata or collaborations. ○ For big social data, this includes embedded metadata. ○ With big social data, it is much harder to collaborate with data creators	Big social data ● The out-of-context effect is amplified by the large scale of the data and the researchers' lack of knowledge about the research subjects.
Data quality and trust-worth-iness	● For both types of data, researchers need to be able to trust reused or collected data. ● For both types of data, researchers can provide information about data quality via metadata and documentation.	Qualitative data reuse ● Research subjects, reporters or recorders of field data, researchers, and data coders can all introduce errors via simple mistakes/inaccuracies, or systematic bias. Big social data ● More subject to self-performance. ● Distortion, errors, and bias from fake accounts and bots. ● Representative sampling issues: users of social media may not be "complete" or representative of society as a whole.

Data comp-arability	For both types data researchers should: ● (1) identify the extent of missing data; ● (2) identify convergence of primary and secondary research questions; ● (3) assess the methods used to produce the primary data	Big social data ● Metadata interoperability may be an issue when trying to combine data from different social media platforms.
Inform-ed consent	● Even if participants consent to archiving, the variety of potential future uses prevent fully informed consent.	Qualitative data reuse ● Consent for the initial study may not extend to data archiving and reuse. ● In the case of deidentified data, participants cannot be contacted to obtain informed consent for new research. Big social data: ● Some terms of service may contain blanket consent agreements that include consent to research. ● Most users do not read the terms of service carefully. ● Scale of data may make it difficult or impossible to obtain informed consent.
Privacy and confid-entiality	● Deidentification may compromise integrity, quality, and context ● Deidentification is not guaranteed to prevent deductive disclosure ● Restricted access can support data reuse if deidentification is not possible or desirable.	Big social data ● Issue of public vs private: posts may be "public," but users may intend for their posts to speak to a personal/private online community. ● Full-text searching and size of datasets make deidentification of big social data difficult.
Intellect-ual property	● Both types of data are the intellectual property of the participants. ● Participants may waive their copyright. ● A licensing agreement can outline the rights, responsibilities, and obligations of researchers. ● Fair use doctrine can be applied.	Big social data ● It is difficult to contact participants to negotiate waiving copyright or licensing. ● Social media companies implement terms of service that govern the behavior of users, developers, researchers, and archivists

Table 7. Key issues and data curation strategies

Issue	Qualitative data reuse	Big social research	Data curation strategies
Context	Data may not be able to be properly understood outside of their original context without the knowledge/expertise of original researchers.	Big social data are taken from a broader context of personal and public life. Out-of-context effect is compounded when data are amassed on a large scale.	● Documentation ● Archiving and sharing related data ● Guidance for balancing context with participant privacy
Data quality and trust-worth-iness	Reusers of archived qualitative data must put their trust in the researchers who designed the study and co-created the data with the research subjects. Errors may be introduced via simple mistakes/ inaccuracies, or systematic bias.	Big social data may be more subject to self-performance. Distortion, errors, and bias may result from fake accounts and bots. Representative sampling issues—social media data may not be complete or representative.	● Documentation ● Trustworthy repositories ● Combining datasets
Data comp-arability	Issues regarding convergence of primary and secondary research questions. Issues regarding aligning methods used to produce the primary data with new research methods.	Metadata interoperability issues when trying to combine data from different social media platforms.	● Documentation ● Metadata standards
Inform-ed consent	Even if participants consent to data sharing, the variety of potential future uses prevents fully informed consent. Consent for the initial study may not extend to data archiving and reuse. In the case of deidentified data, participants cannot be contacted to reconsent to new research.	Some social media platform terms of service may contain blanket consent agreements that include consent to research. Most users do not read the terms of service. Scale of data may make it difficult or impossible to obtain truly informed consent.	● Broad consent ● Alternative consent strategies ● Guidance for risk-benefit analyses
Privacy and confid-entiality	Deidentification may compromise quality and context, and is not guaranteed to prevent deductive disclosure.	Public vs private: posts may be “public,” but users may intend for their posts to speak to a personal/private online community. Full text searching and size of datasets make	● Deidentification procedures ● Restricted access ● Data use agreements ● Guidance for risk-benefit analyses

		deidentification of big social data difficult.	
Intellect-ual property	Data are the intellectual property of participants. Participants may waive their copyright or may license data for use. Fair use doctrine may apply.	Difficult to contact participants to negotiate waiving copyright or licensing. Social media companies implement terms of service that govern the behavior of users, developers, researchers, and archivists.	● Rights management ● Data licensing ● Alternative archiving strategies

4.4. Chapter summary

By investigating issues in qualitative data reuse and big social research and comparing them side by side, data curation practices can be developed to support sounder practices for both qualitative data and big social data. The issues synthesized and the questions outlined here will be answered in the remainder of this dissertation, through semi-structured interviews with researchers and data curators. Chapter 5 describes my research methods. Chapter 6 summarizes the results of the semi-structured interviews, and Chapter 7 provides a discussion of key takeaways.

Chapter 5. Research design

In this chapter, I describe the theoretical framework underlying my research, and I outline my methodological approach. I then describe in detail how I put those research methods into practice. This dissertation is situated in a social constructivist paradigm, which emphasizes how cognitive processes and social environmental factors lead to knowledge formation. The work also incorporates the ideas of communities of practice and epistemic cultures—two theories that help social science researchers group and analyze scientific communities. My research was conducted using a two-part process: a review of the literature to inductively identify key themes, and semi-structured interviews of qualitative researchers, big social researchers, and data curators to further investigate and expand upon those themes.

5.1. Theoretical framework

Information science explores multidisciplinary issues, with an ultimate aim of understanding how people interact with information (Bates, 1999). Contemporary information science research has been built upon what Murray and Evers (1989) refer to as "theory borrowing"— the practice of building upon theories from multiple disciplines including social science and the humanities (Pettigrew & McKechnie, 2001). Cronin suggests that the past few decades have brought about a "sociological turn" in information science research, built on a foundation of social constructivism (2008, p. 471). Social constructivism, also called "collectivism" by Talja, Tuominen, and Savolainen (2005), is based on Vygotsky's social constructivist theory of cognitive development, which emphasizes that knowledge formation derives from a combination of cognitive processes and social environmental factors:

> Knowledge formation and the development of knowledge structures take place within a socio-cultural context. Individual development derives from social interactions within which cultural meanings are shared by a group and eventually internalised by the individual. It is assumed that individuals construct knowledge in interaction with the environment and that in the process both the individual and the environment are changed. Thus, the subject of study is the dialectical relationship between the individual and the socio-cultural milieu. (Talja et al., 2005, p. 85)

Theories built upon the social constructivist paradigm are commonly used in information science research. Some examples include the ideas of social and cultural capital (Bourdieu, 1986), the theory of the network society (Castells, 2000), ethnomethodology (H. Garfinkel, 1967), diffusion of innovations theory (Rogers, 2003) and actor-network theory (Latour, 1996).

By framing my research with social constructivist theory, my dissertation aims to synthesize the insights and approaches of qualitative researchers, big social researchers, and data curators to support ethical, legal, and epistemologically sound data sharing. To further my goal of understanding of the communities investigated in my dissertation (qualitative and big social research communities, and the data curation community), I also incorporate the ideas of communities of practice (Lave & Wenger, 1991; Wenger, 1998) and epistemic cultures (Knorr-Cetina, 1999)—theories that help social science researchers group and analyze scientific communities.

Communities of practice are "groups of people who share a concern, set of problems, or a passion about a topic, and who deepen their knowledge and expertise in this area by interacting on an ongoing basis" (Wenger et al., 2002, p. 4) The goal of communities of practice theory is to explain *how* groups of *people disseminate knowledge.* The groups I am examining in this dissertation are qualitative researchers who reuse or archive data, big social researchers, and data curators. Each of these communities of practice has three key characteristics: their domain, their community, and their practice (Wenger et al., 2002). For the purposes of this dissertation, the domains are the interests and disciplines relating to the practices of qualitative data reuse, big social research, and data curation; the communities form when researchers or curators work together, discuss, and share the interests and disciplines that characterize their domain; practice includes the shared research practices, shared jargon, and shared values of each community. Communities of practice theory has been used to study science collaboratories (Bos et al., 2007), to build data management and digital scholarship services in academic libraries (P. L. Smith et al., 2020), and as a framework for integrating educational research and practice (Buysse et al., 2003).

The idea of epistemic cultures is focused on the *processes of creating knowledge*; consequently, epistemic cultures not only include people and groups of people, but also the objects and technologies they use to discover or develop knowledge. For this dissertation, engaging with epistemic cultures theory means considering qualitative data reusers, big social researchers, and data curators as researchers and practitioners, and also considering the tools they use to communicate with their research teams and respondents, to collect and analyze data, and to curate data. The theory of epistemic cultures has been used to compare disciplinary approaches (Heidler, 2017; Stevens et al., 2020) and to develop pedagogical approaches (Michel & Tappenbeck, 2019). As Borgman writes: "Common to both communities of practice and epistemic cultures is the idea that knowledge is situated and local. Nancy Van House (2004) summarizes this perspective succinctly: 'There is no 'view from nowhere'—knowledge is always situated in a place, time, conditions, practices, and understandings. There is no single knowledge, but multiple knowledges'" (Borgman, 2012, p. 1062).

5.2. Methodology

My ultimate aim in conducting this research is to understand how the ideas and approaches of two distinct research communities—qualitative researchers and big social researchers—can be combined so that a third community—data curators—can develop and encourage stronger data curation practices, thus leading to more ethical, legal, and epistemologically sound data sharing.

Accomplishing this aim requires an in-depth understanding of researchers' behaviors and attitudes. Such in-depth understanding can be facilitated by a qualitative approach, rooted in grounded theory (Glaser & Strauss, 1967), to iteratively produce insights. My research followed a five-stage process, using the critical incident technique.

Critical incident technique is a widely used, established practice in qualitative research (Butterfield et al., 2005), and has been used extensively in information science literature, including in user studies, as a method for understanding library systems and

information-seeking behaviors, and to study library human resources and management (Lipu et al., 2007; Wildemuth, 2017). My research especially follows the lead of Faniel, Kriesberg, and Yakel (2016), who use critical incident technique to understand data reuse, and Cushing and Dumbleton (2017), who use this technique to investigate personal information management behaviors.

An interview that uses critical incident technique is structured around a specific incident, defined by Flanagan as "any observable human activity that is sufficiently complete in itself to permit inferences and predictions to be made about the person performing the act" (1954, p. 327). The "incident" in critical incident technique is identified as a specific example that focuses a participant's answers to the interview questions. This focus allows participants to remember more detail and provide concrete examples and experiences (Wildemuth, 2017). Critical incident technique follows a five-stage process: (1) determine the general aims of the activity to be studied; (2) set specifications for data collection, including the types of situations to be observed or reported and the incident's relevance and effect on the general aim of the activity; (3) collect data via interviews or questionnaires centered around relevant incidents; (4) analyze the data; and (5) interpret and report the findings (Borgen et al., 2008).

I accomplished Stage 1 of this process (determining the general aims of the activity to be studied) by conducting the literature review described in section 5.3. Stage 2 (establishing the specifications for data collection), is outlined in Table 8 below. This table identifies the activity, aims, situation, critical incidents, and critical behaviors relating to my research questions.

Table 8. Critical incident technique, Stage 2. Specifications for data collection (adapted from Hughes et al., 2007)

Specification	Explanation
Activity	Curating qualitative data and big social data.
Aim of the activity	Understanding how data curation practices can support epistemologically sound, ethical, and legal qualitative data reuse and big data research.
The situation	Who? Big social researchers, qualitative researchers, and data curators. Where? The United States. What? Considering the epistemological, ethical, and legal issues that arise when using big social data, reusing qualitative data, or curating and sharing qualitative or big social data.
Critical incidents	Experiences involving either the use of big social data, the reuse of qualitative data, or the curation and sharing of qualitative or big social data, especially regarding the six key epistemological, ethical, and legal issues—context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property.
Critical behaviors	Instances and actions involving the use and reuse of big social data, the reuse of qualitative data, or the curation and sharing of qualitative or big social data, with a focus on each of the six key epistemological, ethical, and legal issues—context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership.

In Stage 3 of the research process, I collected data using semi-structured interviews that centered around specific incidents of qualitative data archiving or reuse, big social research, or data curation. Semi-structured interviews have been used to study data sharing behaviors and attitudes (e.g., Faniel et al., 2019; Faniel & Connaway, 2018; Yoon, 2017; Zimmerman, 2008) and to study the behaviors and attitudes of communities of practice and epistemic cultures (Ardichvili et al., 2006; Keller & Poferl, 2016). Semi structured interviews are described in more detail in section 5.4. In Stage 4 of the research, I analyzed the collected data. A key assumption of qualitative research is that "the world is neither stable nor uniform, and therefore, there are many truths" (Bloomberg & Volpe, 2016, p. 186); with this in mind, I used both inductive and deductive analysis approaches, allowing for some flexibility in research design to support the iterative development of insights. In Stage 5 of the research, I interpreted my results and reported my findings (see Chapter 6, Results).

5.3. Literature review

Using the methods outlined by Creswell (2009) and detailed in the *Handbook of research synthesis and meta-analysis* (H. M. Cooper et al., 2019), I conducted an inductive research synthesis of the literature on qualitative data reuse and big social data research. According to Cooper et al., "research syntheses…pay attention to relevant theories, critically analyze the research they cover, try to resolve conflicts in the literature, and attempt to identify central issues for future research" (2019, p. 6). My research synthesis consisted of the following steps: literature search, data evaluation, data analysis, and interpretation of results (H. M. Cooper et al., 2019).

For the literature search, I searched the library catalog and online databases using the following strings:

- "qualitative secondary analysis"
- "qualitative data reuse"
- "qualitative data archiving"
- "social media data"
- "social media data archiving"
- "big social data"

While reviewing initial articles, I identified further reading among the cited works (C. Cooper et al., 2017), also called citation chaining (Hu et al., 2011). This search process yielded approximately 300 articles. I coded each article according to my key themes, inductively identifying these themes as I read more articles. My coding focused on (1) research objectives and methods; (2) discussions of theory, including epistemological and ethical issues; and (3) data curation practices. The themes that emerged during this initial stage were consent, methodology, privacy/confidentiality, context, trust/data quality, metadata/transparency, archiving, restriction/data access considerations, intellectual property/data ownership; data value, and data credit/citation. Six central issues emerged in common between qualitative data reuse and big social data research—context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership.

After these six issues emerged, I continued identifying and reading articles in each area, pinpointing and continuing to examine specific areas. I describe each issue and related sub-issues in detail in Chapters 2 and 3 of this dissertation, and then I synthesize how these issues relate to both qualitative data and big social data in Chapter 4. I used the six key issues—context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership—to structure three interview guides for the semi-structured interviews I conducted. The semi-structured interview process is described further below.

5.4. Semi-structured interviews

The six central issues that I identified through my research synthesis informed the second phase of my research—semi-structured interviews with three different types of participants, referred to throughout this study as "communities of practice:"

- researchers who have used big social data
- qualitative researchers who have published or reused qualitative data
- data curators who have worked with one or both types of data

As stated above, my research aim is to understand the data curation implications of the similarities and differences between big social research and qualitative data reuse. The three communities of practice discussed in this dissertation can identify and speak to the similarities and differences in each area, thus supporting, rejecting, or adding nuance to the tentative conclusions from the literature review I conducted in Step 1 (see section 5.3).

Semi-structured interviews are particularly useful when investigating research questions in which "the concepts and relationships [at issue] are relatively well understood" (Ayres, 2008, p. 811)**.** This method is commonly used in grounded theory research because the researcher has "more direct control over the construction of data than does a researcher using most other methods, such as ethnography or textual analysis" (Charmaz, 2001, p. 676), thus giving researchers more analytic control over the data through the flexibility of ended questions; as

new ideas continually emerge throughout the interview process, the interviewer has the flexibility to pursue these new ideas (Charmaz, 2001).

Semi-structured interviews are used to collect data from participants who are knowledgeable in relevant areas. Expert interviews are widely used in social science research as a strategy for reaching participants who can contribute key insights. Some scholars have questioned how expert participants are identified, what features or characteristics define an expert, and how different types of expert interviews fit into different research designs (Bogner et al., 2009)**.** For the research described in this dissertation, the interviews were designed to gather data about the participants' personal practices when conducting big social research, qualitative data reuse/sharing, and data curation. I consider researchers and data curators to be experts in their own work, and therefore valuable informants. Because I wanted my study to encompass a broad sample of researchers, I aimed for as much heterogeneity among my informants as possible—looking for researchers who were at various stages of their careers and were trained in various disciplines (Schreier, 2018) (see Table 11. Qualitative researchers by discipline, Table 12. Big social researchers by discipline, and Table 13. Number of participants by rank or role).

The interviews were designed to prompt in-depth discussions of the situations in which key issues arise (Wildemuth, 2017), using critical incident technique to prompt participants to identify one situation that would serve as a springboard for discussing the specifics of key issues. I selected semi-structured interviews as an appropriate research method for three reasons. First, the literature review showed that data curation for big social research and qualitative data reuse is conducted by both researchers and data curators, so I wanted to speak to experts from each of these communities. Second, big social research and qualitative data reuse are both emerging domains, with evolving ideas; interviewing experts would therefore add value to the literature review in Step 1. Third, the three communities of practice (big social researchers, qualitative data reusers, and data curators) would each have varying perspectives on the key issues, thus providing new insights from each community about different challenges they might face and their potential strategies for addressing these challenges.

This semi-structured interview process was informed by the guidelines for information science researchers laid out by Luo and Wildemuth (2017). To conduct the interviews, I developed three interview guides—one for qualitative researchers, one for big social researchers, and one for data curators. Adhering to the approach described by Luo and Wildemuth, these guides included "essential questions, extra questions, throw-away questions, and probing questions" (2017, p. 234). The questions were centered around the six key issues identified during my literature review—context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership.

I submitted my research plan to the Montana State University Institutional Review Board (IRB), which approved the interview guides and the consent agreement, and which designated my research as exempt under MSU IRB Exempt Protocol #SM022421-EX. (See Appendix 1 for the initial consent agreement and Appendix 2 for a minorly modified consent agreement.)

5.4.1. Developing the interview guides

To develop the three interview guides, I initially began by outlining hypotheses for each of the six key issues. Then I created questions that would test each hypothesis. These hypotheses are listed in Table 9, organized by issue.

Table 9. Hypotheses by issue

Issue	Hypothesis
Context	1. Qualitative researchers have a stronger concern about context than do big social data researchers 2. Context can be communicated to some extent through embedded or added metadata. a. Data curators who specialize in archiving qualitative data can also support metadata that preserves context for big social research and big social data.
Data quality and trustworthiness	1. Big social data is more prone to quality issues than shared/reused qualitative data. 2. Documentation/metadata can support data quality.

Data comparability	1. Comparing and combining data enables higher quality research (e.g., larger scale, more representative samples, broader conclusions). 2. Combining datasets is made more difficult for those who reuse qualitative data or use big social data because of challenges relating to missing data, research questions, methods, and metadata interoperability. 3. Data comparability issues are similar for qualitative data and big social data.
Informed consent	1. Qualitative and big social researchers have different values and considerations regarding informed consent. a. Qualitative researchers are more strict about issues related to consent and reuse, even for archived data/data reuse. b. Big social researchers are more open to using creative strategies to address consent (e.g., focus groups, community advisory groups). 2. Qualitative data curation approaches for consent could be adapted to fit big data researchers.
Privacy and confidentiality	1. Big social data researchers are less concerned about privacy than qualitative researchers. 2. Data curation practices for supporting privacy with qualitative data can inform big social data.
Intellectual property and data ownership	1. IP is a more important issue for big social researchers than researchers who reuse qualitative data. 2. IP concerns may prevent big social data researchers from archiving data. 3. Data curation practices that address IP issues for qualitative data can inform big social data and vice versa.

I pre-tested the three interview guides using a two-part method. First, each guide was reviewed by three experts: (1) Vivien Petras, my dissertation advisor; (2) Kalpana Shankar, a mentor assigned to me during the ASIS&T Doctoral Colloquium; and (3) Eric Raile, the director of the Montana State University HELPS Lab (HELPS Lab, 2020), a service at MSU that provides assistance for human subjects research. Each expert reviewer identified technical problems with the guide and provided suggestions for improvement. This expert review phase helped me to better structure the questions, especially a suggestion from Dr. Shankar that I use critical incident technique (FIanagan, 1954), described in more detail above, in section 5.2. Following Dr. Shankar's suggestion, I reworked the interview questions to align with critical incident technique, asking that participants identify one specific incident—either research with big social data, research that reuses qualitative data, or the curation process for either qualitative or big social data.

This expert review phase helped me refine the interview questions so that they would allow me to better understand what point (if at all) the participants considered the various key issues during their critical incident. Then, if participants did consider an identified issue, the guide would elicit explanations of what resources the participant turned to and what strategies they used to address the issue.

After the expert review phase, I pre-tested the interview guides with two test participants to refine my questions and procedures. This pre-testing revealed parts of some questions that were unclear and identified certain questions that needed more detail or follow-up options. This second round of revision resulted in smaller but important refinements to the interview guides.

Ultimately, I developed three final interview guides—(1) an interview guide for researchers who had shared or reused qualitative data, (2) an interview guide for big social researchers, and (3) an interview guide for data curators. (See Appendix 3, Appendix 4, and Appendix 5 for the full text of the interview guides.) All three guides are structured around the six key issues identified in the literature review—context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. All three guides begin with an introduction of the research question and the main ideas to be explored, and an elicitation of a critical incident that the participant will discuss. Then all three guides ask a warm-up question (Luo & Wildemuth, 2017):

- For qualitative and big social researchers: Tell me about the type of research you do and what kind of data you produce.
- For data curators: Tell me about the types of data you usually curate and what your interests are regarding data curation.

From there, the main interview begins with the following introductory questions about the data involved in the critical incident:

- Please describe your data collection method (API, scraping, shared dataset, etc.)
- Was this example part of a grant-funded project that required specific treatment of the data? E.g., did you have a data management plan?
- Is any of the data from your example published?

- Is the data published in a repository? Which one?
- What are the plans for storing, retaining, and deleting data in the future?
- Who has access to the data?

Sections 5.4.1.1 and 5.4.1.2. describe the specifics of the remainder of each interview guide.

5.4.1.1. Interview guides for qualitative researchers and big social researchers

When interviewing qualitative researchers who had either shared or reused qualitative data, and when interviewing big social researchers, I aimed to elicit (1) specific examples of when the interviewees had encountered challenges with qualitative data reuse, qualitative data sharing, and/or big social research, and (2) what strategies they used to address these challenges. By using critical incident technique, I prompted my interviewees to discuss how each of the six key challenges related to a specific research project they had recently worked on; by continually returning to the critical incident, I aimed to avoid eliciting generalities or platitudes, instead prompting my interviewees to provide me with real-life, specific examples.

To facilitate these goals, in my interview guides for qualitative and big social researchers, each key issue had an initial question prompting the participant to identify a specific time when they considered each key issue:

- “Tell me about a time (if any) during the process of your example when you considered the issue of”... [context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, intellectual property and data ownership]. The guide provided more detail about what is meant by each of these concepts, including potential examples of how these key issues might arise.

The guides then provided several options for follow-up questions about the strategies the participants used to address each issue, including:

- “Did you consult with anyone, consider other research projects, or refer to literature, policies, or guidelines regarding this issue? Please explain.”

The follow-up questions acted as prompts for me as the interviewer. I often asked at least the question quoted above, but the semi-structured interview format allowed for flexibility in my follow-up questions—either adhering to the questions in the guide, revising them slightly, or posing new questions according to my own intuition.

5.4.1.2. Interview guide for data curators

When interviewing data curators, I used much the same approach as I did in my interviews with qualitative and big social researchers. By using critical incident technique, I aimed to elicit discussion of specific examples of the six key issues, as well as the strategies the data curators used to address those issues. Additionally, I hoped that data curators would have a broader perspective on how the other two communities of practice (i.e., qualitative data reusers/sharers and big social researchers) related to one another.

In the interview guide for data curators, each key issue had an initial question prompting the participant to identify a specific time when they considered each key issue:

- "During your example, what challenges did you encounter (if any) relating to"... ... [context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, intellectual property and data ownership].

The guide then provided options for follow-up questions that asked participants to describe any strategies they used to address these issues, including:

- What strategies did you use to communicate, describe, or clarify the challenges in your example?

As with the guides for qualitative and big social researchers, the follow-up questions in the guide for data curators acted as prompts for me as the interviewer, and the semi-structured interview format gave me flexibility in my follow-up questions—either adhering to the questions in the guide, revising them slightly, or posing new questions according to my own intuition.

A final follow-up question in the guide for data curators was:

- "(If applicable) What similarities and differences do you see between data curation strategies that address issues of [context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, intellectual property and data ownership], for qualitative data and big social data?

I posed this question under the assumption that data curators may have a better awareness or understanding of any substantive similarities and differences between these two types of data, and may have developed data curation ideas and strategies in response to the similarities and differences they had encountered.

5.4.2. Sampling

I used a few different strategies to select my interview participants, depending on participant type.

For data curators, I identified participants through my literature review, contacting authors of key articles. I also used my knowledge of the data curation community to contact data curators who I knew had experience with qualitative and big social data.

For qualitative researchers, I used two strategies to select participants. To identify participants who had published qualitative data, I searched the Qualitative Data Repository (Center for Qualitative and Multi-Method Inquiry, 2020) for datasets that had been published in the last four years; I also searched Dryad (Dryad, 2022) and Zenodo (CERN Data Centre, 2020) for qualitative data, using the keywords "interview" and "qualitative." To identify researchers who had reused qualitative data, I searched the Web of Science database for the keywords "qualitative data reuse" and "qualitative secondary analysis," then filtered for articles published in the past four years. I was only able to connect with two researchers who had reused data from sources that were not their own research. This is indicative of the current rarity of qualitative data reuse. The final group of qualitative researchers included two researchers who had reused qualitative data from other sources, three researchers who had conducted secondary analysis on their own qualitative data, and five researchers who had shared their own qualitative data in a repository.

For big social researchers, I searched Web of Science for the keywords “big social data,” “social media data,” “social media,” “facebook,” “twitter,” “reddit,” and “pinterest,” then filtered for articles published in the past four years.

Additional interviewees were identified by asking my dissertation advisors and mentors for suggestions.

After I identified this initial group of participants, I added participants using snowball sampling. Snowball sampling—also called chain referral or network sampling—is an established method for augmenting a participant list, first developed in the 1960s (Kadushin, 1968). This method uses an initial list of key participants as “seeds” who then offer suggestions, from the perspective of the participant community, about who else should be interviewed on this topic. This sampling method is often used when interviewing potential participants who come from a relatively small professional population and who are therefore likely to be connected to each other (Bernard et al., 2017). Thus, snowball sampling is an appropriate method for this dissertation, which focuses on communities of practice who are conducting specialized research, data sharing, and curation activities.

In addition to snowball sampling, I used theoretical sampling—that is, responsive sampling conducted at the same time as my interviewing and data analysis. By using theoretical sampling, I was able to selectively identify potential participants according to the concepts I had derived from my analysis and any questions or gaps I identified along the way (Corbin & Strauss, 2008). For instance, I had a higher initial response rate from qualitative researchers who had published data in a repository, and I noted a gap in my analysis regarding the viewpoints of participants who had reused qualitative data; I therefore purposefully searched Web of Science for additional participants who had reused qualitative data. I continued my sampling until I reached saturation—that is, “the point in the research when all the concepts are well defined and explained” (Corbin & Strauss, 2008, p. 145).

I limited my participants to those working in the United States. By limiting my context to one country, I aimed to eliminate potential challenges and complexities due to differences in

laws, policies, and infrastructure (Chawinga & Zinn, 2019; Mulder et al., 2020; Tenopir et al., 2011).

The positive response rate to my requests for interviews varied by type of participant. Data curators had the highest percentage of positive responses, with a 55.6% positive response rate. Qualitative researchers had a 37% positive response rate, and big social researchers had the lowest positive response rate, at 15.4%. Table 10 provides more information on response rates by type of participant.

Table 10. Response rates by type of participant

Type of participant	Interview requests	Positive responses	% positive response
Data curators	18	10	55.6%
Qualitative researchers	27	10	37%
Big social researchers	65	10	15.4%

The 20 qualitative researchers and big social researchers whom I interviewed came from a variety of disciplines. Information Science is somewhat over-represented in my dataset because Information Science researchers were more likely to respond positively to my interview requests. This may be due to their interest in the research topic or their knowledge of my dissertation advisors. Because Information Science leans toward interdisciplinarity (Chang, 2018), the different critical incidents discussed by Information Science researchers were distinct enough that the sample still provides a broad variety of disciplinary ideas. Tables 11 and 12 provide overviews of qualitative and big social researchers by discipline.

Table 11. Qualitative researchers by discipline

Discipline	Number of participants
Information Science	4
Anthropology	2
Public Health	1
Education	1
Nursing	1
Social Work	1

Table 12. Big social researchers by discipline

Discipline	Number of participants
Civil Engineering	2
Communication	2
Computer Science	2
Information Science	2
Journalism	1
Public Health	1

Participants came from a variety of ranks and roles. Data Curators were most represented, with six participants who were curators at repositories. The dataset also has high representation among Assistant Professors, Post Doctoral Scholars, and Academic Librarians. Table 13 provides an overview of the number of participants from each rank or role.

Table 13. Number of participants by rank or role

Rank or Role	Number of participants
Data Curator	6
Assistant Professor	5
Post Doctoral Scholar	4
Academic Librarian	4
Associate Professor	3
Professor	3
Research Scientist	2
PhD Student	1
Professional Staff	1
Non-Tenure Track Faculty	1

5.4.3. Interview process

I interviewed ten participants from each of the three target populations—big social researchers, qualitative researchers who had published or reused data, and data curators. The interviews were conducted between March 11 and October 6, 2021. The longest interview lasted one hour and 13 minutes and the shortest lasted 33 minutes, with an average interview length of 53 minutes. (See Appendix 6 for exact interview dates and lengths.)

All interviews were scheduled using Microsoft Bookings software (Ako-Adjei & Penna, 2021) and conducted using Zoom videoconferencing software (Zoom, 2021). When I scheduled the interviews, I emailed the participants to ask them to identify a critical incident prior to the interview, as well as to provide them with the IRB-approved consent agreement for their review and the full text of the applicable interview guide, which included a short description of the research. (See Appendix 7, Appendix 8, and Appendix 9 for the emails sent to participants) At the beginning of each interview, I emailed the participant the consent agreement using DocuSign, a system that facilitates signatures for official documents

(DocuSign, 2021). After the participant had received the agreement via DocuSign, I reviewed each section of the consent agreement with the participant. I asked if the participant had any questions (no participant asked any questions), and then we each electronically signed the consent agreement using DocuSign.

After conducting the first 10 of my 30 interviews, I asked the Montana State University IRB for a minor modification of my consent agreement, because, after speaking with several participants, I realized that I wanted participants to be able to opt in or opt out of allowing the deidentified transcript from their interview to be published in a data repository. I therefore obtained approval to add a check box at the bottom of the consent agreement to support this option[2]. I also noticed that some participants viewed discussing their work practices as a potential risk; I therefore added the following language to the statement regarding risk: "There are no major risks to participating in the study, but you will be discussing issues you encounter in your work and research practices." (For the initial and revised versions of the consent agreement, please see Appendix 1 and Appendix 2.)

At the beginning of each interview, I introduced myself and gave an overview of the research I was conducting—reading from the description that was provided on the first page of the interview guide. I then explained that there would be eight question areas—an introductory section, one section for each of the six key issues identified in my literature review, and a wrap-up section. I asked for permission to record the interview, which I did using the built-in recording technology of Zoom videoconferencing software (Zoom, 2021). I also took notes during the interviews.

In the semi-structured interviews that I conducted, a variety of different questions, follow-up questions, prompts, and topics informed the outcomes of the interviews, but the semi-structured format gave me flexibility to change course, following my curiosity and intuition, to expand upon the research questions and further the goals of the interview (Durdella, 2019). As dictated by my semi-structured interview structure, my questions

[2] In the interviews conducted after implementation of the data sharing opt-out option, one participant opted not to share their data, and one participant asked to review their deidentified transcript prior to data publication.

generally followed the order of the interview guide, but sometimes bounced around between key issues according to the trajectory of our conversation. I asked some follow-up questions that I had anticipated in the interview guide, as well as other follow-up questions that occurred to me in the moment and were specific to the conversation at hand. Most interviews went smoothly, although I note some limitations in Chapter 8, section 8.2.

After each interview, I verbally thanked the participant. I also sent a follow-up email to each participant, thanking them for their time and their insights.

5.5. Analysis

I used Otter.ai speech-to-text software (Otter.ai, 2021) to create initial transcriptions of the interview recordings. I hired an undergraduate student to hand-edit the transcripts for accuracy. The student made notes when they had questions or when the recording was unclear, and I conducted a final review of the transcripts for accuracy.

I also conducted an initial deidentification of the transcripts at this stage, in the summer and fall of 2021. Additional deidentification was conducted in partnership with the curators at the Qualitative Data Repository, where the transcripts will be shared in late 2022 (Mannheimer, 2022). If you would like to access the data prior to late 2022, special permission will be provided by request. (Please see Appendix 13 for full data availability information.)

5.5.1. Coding

I analyzed the interview transcripts using a qualitative content analysis of the interview transcripts. This involved using a combination of inductive and deductive coding approaches, as outlined in Zhang and Wildemuth (2017) and as detailed in Bernard, Wutich, and Ryan (2017). After reviewing the research questions, I used NVivo software to identify chunks of text in the interview transcripts that represented key themes of the research (QSR International, 2022). Because the interviews were structured around each of the six key issues that I had identified in the literature review (see Chapters 2 and 3), I deductively created a parent code for each of the six key issues—context, data quality and

trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. I then used inductive coding to create subcodes beneath each of the parent codes for these key issues.

As I continued, I normalized the themes by comparing any new themes to previous themes, in accordance with grounded theory's constant comparative method, an iterative process of reading and analyzing text to reveal themes (Glaser and Strauss, 1967). This iterative process includes comparing data with other data, coding data with initial codes, identifying focused codes, comparing and sorting coded data, grouping the codes into broad categories, comparing data and codes with these broad categories, constructing theoretical concepts from categories, comparing category with concept, and comparing concept with concept (Charmaz, 2008).

My initial coding resulted in 412 different codes (see Appendix 10). Most codes were subcodes of one of the six issues, but some codes also emerged that suggested additional parent codes. At this stage, I had additional parent codes for methodology, curation, data sharing, disciplinary ideas, and synthesis. Many codes at the initial stage were closely related to one another, but with different levels of granularity. For instance, I had created the following three codes:

1. "consent - asking permission for direct quotes,"
2. "consent - quoting tweets didn't feel right," and
3. "consent - taking care with direct quotes."

The first and second codes include specifics about what type of care was taken when using direct quotes. Several times throughout the coding process, I reviewed the codes and combined codes such as these three codes addressing direct quotes. To keep a record of my progress at this stage, instead of immediately merging the codes, I would move related codes so that they were subcodes of the broadest concept. NVivo software allows the user to drag and drop subcodes to nest below parent codes. In the example above, I dragged the first and second codes—"consent - asking permission for direct quotes" and "consent - quoting tweets didn't feel right," dragging them to nest underneath the third and broadest code—"consent - taking care with direct quotes."

In October 2021, about half-way through the coding process, I conducted a pile sorting exercise as outlined in Bernard, Wutich, and Ryan (2016)—although instead of using slips of paper, I copied and pasted themes into a spreadsheet, sorting into columns rather than piles. This initial sorting exercise helped me begin to discern the patterns in the quotes and themes that I had identified. However, the categories during this first sorting exercise ended up being very broad. I had identified categories such as "challenges," "strategies," and "tools." While this exercise helped me understand my themes better, I ultimately disregarded this spreadsheet because I considered it too broad to be useful. I needed the codes to have analytical usefulness—that is, I needed the codes to be able to help me answer my research questions and support new insights into qualitative data reuse and big social research.

To corroborate and potentially enhance the analytical usefulness of the codes, in November and December 2021, I used NVivo to conduct a second pile sorting exercise. This time, I simply continued the strategy of dragging and dropping codes to become subcodes of broader codes. During this pile sort, I grouped all of the codes so that no codes had only one reference in the interviews. This strategy also gave me an opportunity to conduct a second review of my sorting thought-process. In late December 2021 and January 2022, after I had nested the codes to my satisfaction, I reviewed each one to ensure that the subcodes fit into the broad code I had selected. I then used the NVivo function "Merge into Selected Code Removing Original," permanently combining the more granular codes into the broader, more analytically useful code. The final pile sort resulted in nine parent codes and 104 subcodes (see Table 14). (See Appendix 11 for the full final codebook).

Table 14. Parent codes and related number of subcodes

Parent code and definition	Number of subcodes
Context - Maintaining and understanding context of reused data	21 subcodes
Quality - Data quality and trustworthiness, trust in data creators	12 subcodes
Comparability - Data comparability and interoperability	5 subcodes
Consent - Participant informed consent	22 subcodes
Privacy - Privacy and confidentiality of data	14 subcodes
Intellectual property - Intellectual property and data ownership issues	11 subcodes
Domain differences - Differences between big social researchers and qualitative researchers, and how these differences affect practices	4 subcodes
Strategies for responsible practice - Strategies used by participants to support ethical, legal, and epistemologically sound research.	5 subcodes
Data curation issues - Issues relating to data curation theory and practice, including benefits, challenges, and complexities	10 subcodes

5.5.1.1. Context

This category refers to the idea of maintaining, explaining, and understanding the original context of reused qualitative data and big social data. This category includes quotes that discuss the challenges of context, such as the tension between successful deidentification and the preservation context. It also includes ideas about strategies for supporting and maintaining context, such as documentation and research design. This category had a total of 21 subcodes. Subcodes with the highest number of mentions were:

- context - description, metadata, documentation to support context (n=13)
- context - in tension with privacy (n=10)
- context - good documentation is time consuming (n=7)
- context - including related materials with data (n=9)
- context - big social data - interface and features provides context (n=6)
- context - may be difficult to ascertain with big social research (n=7)

- context - representativeness of data (n=5)

5.5.1.2. Quality

This category refers to the idea of communicating the level of data quality and trustworthiness when sharing data, and assessing the quality of reused data and big social data. This category includes quotes that discuss potential quality-related pitfalls such as spam and bots, missing data, and trust in data creators. It also includes ideas about strategies to support or maintain data quality, such as in-depth documentation and metadata. This category had a total of 15 subcodes. Subcodes with the highest number of mentions were:

- quality - description, metadata, documentation support data quality (n=18)
- quality - data completeness (n=10)
- quality - issues with large-scale and automated collection (n=7)
- quality - spam and bots (n=6)
- quality - representativeness of data (n=5)
- quality - curator review (n=4)

5.5.1.3. Comparability

This category refers to the idea of data comparability and interoperability as they relate to qualitative data sharing/reuse, big social data, and data curation. This category includes codes that discuss strategies to support comparability and interoperability, as well as codes that refer to key challenges to comparability and interoperability. This category had a total of six subcodes. Subcodes with the highest number of mentions were:

- comparability - interoperability - formats, metadata, language, etc. (n=11)
- comparability - more data = stronger conclusions (n=10)
- comparability - documentation and metadata (n=9)

5.5.1.4. Consent

This category refers to the idea of informed consent in qualitative data sharing/reuse, big social data, and data curation. This category includes codes that discuss strategies to support informed consent, as well as various ideas as to how consent relates to data sharing/reuse

and big social research. This category had a total of 22 subcodes. Subcodes with the highest number of mentions were:

- consent - IRB (n=22)
- consent - consent language and procedures (n=15)
- consent - public vs. private (n=8)
- consent - taking care with direct quotes (n=7)
- consent - social media terms of service include consent (n=6)
- consent - sensitivity of data (n=5)
- consent - don't know what future uses might be (n=5)

5.5.1.5. Privacy

This category refers to ideas about privacy and confidentiality for reused data and big social data. This category includes quotes that discuss challenges about participant privacy. It also includes ideas about strategies for supporting privacy, such as deidentification, privacy-focused research design, and weighing benefits of data sharing and big social data research against potential harms to participants. This category had a total of 14 subcodes. Subcodes with the highest number of mentions were:

- privacy - deidentification (n=18)
- privacy - restricted access (n=11)
- privacy - sensitivity of data (n=11)
- privacy - considering potential harms (n=10)
- privacy - participant expectations (n=10)
- privacy - research design (n=8)

5.5.1.6. Intellectual property

This category refers to the idea of intellectual property and data ownership. This category includes quotes that discuss participants' challenges and concerns relating to data licensing, ownership, and social media terms of service. This category had a total of 11 subcodes. Subcodes with the highest number of mentions were:

- IP - platform or data provider terms of service (n=13)
- IP - purchasing or using commercially-available data (n=8)

- IP - data sovereignty and ownership (n=7)
- IP - data licensing (n=6)
- IP - lack of clarity about IP laws (n=5)
- IP - data citation (n=5)

5.5.1.7. Domain differences

This category refers to the differences between the communities of practice that I spoke to in my interviews. “Domain” is a term used by Wenger et al. (2002) in their theory of communities of practice; I use this term to describe the combination of interests and disciplines that are present within the communities of practice investigated in this research. This category includes quotes that discuss specific practices, as well as quotes that compare and contrast communities of practice. This category had a total of four subcodes. Subcodes with the highest number of mentions were:

- domain differences - data sharing values and norms (n=12)
- domain differences - research practices and standards (n=9)
- domain differences - skills, training, and background (n=8)

5.5.1.8. Strategies for responsible practice

This category refers to strategies that support ethical, legal, and epistemologically sound qualitative data sharing/reuse and big social research. This category’s quotes mostly relate to concrete strategies that researchers have used, such as discussions with colleagues, conducting formal or informal risk-benefit analyses, reading the literature, and developing research questions that are conducive to responsible practice. This category had a total of five subcodes. Subcodes with the highest number of mentions were:

- strategies for responsible practice - risk-benefit analysis (n=17)
- strategies for responsible practice - discussions with colleagues and collaborators (n=13)
- strategies for responsible research - appropriate research questions and scope (n=5)

5.5.1.9. Data curation issues

This category relates to issues in data curation and data sharing. This category includes quotes that discuss the benefits, challenges, and complexities relating to repository practices, data sharing, and data management planning. This category had a total of 10 subcodes. Subcodes with the highest number of mentions were:

- curation - value of big social research and qual data sharing (n=11)
- curation - cost and time (n=10)
- curation - collaborating with curators and repositories (n=7)
- curation - planning for data sharing makes it less of a hurdle (n=5)
- curation - for transparency (n=4)
- curation - technical requirements of big social data and data reuse (n=4)

5.5.2. Memo writing

Memo writing is an important method in grounded theory that supports connection-making and theory-building; memo-writing helps the researcher think more concretely about the data in order to synthesize ideas and develop key takeaways (Charmaz, 2008). Immediately following each interview, I created a short memo, jotting down a few key takeaways from the conversation. (See Appendix 12 for memos.) I supplemented these memos during the coding process, creating field notes that recorded my "observations, hunches, and insights on the fly" (Bernard et al., 2017, p. 228) during the process. These memos helped me synthesize my ideas about the research. For instance, after I interviewed one big social researcher (BSR03), I was especially struck by their discussion of how they specifically designed research questions in a way that took into account ethical considerations. This researcher specifically opted not to pursue an idea to incorporate research questions about differences between gender identities, because it would have required them to connect their big social dataset with other demographic datasets, potentially leading to identification of participants and potential harm to vulnerable populations. This interview prompted me to write Memo BSR03 about how researchers can choose to ask questions that support more ethical big social data collection, and conversely, they can purposefully choose not to ask research questions that could be problematic. Additional memos that I wrote during the coding process helped me to understand which codes were most important to theory-building. For

instance, a key group of codes discuss the idea of weighing one idea against another in a risk-benefit analysis. In Memo BSR05, I attempt to clarify my ideas about when and why researchers conduct formal and informal risk-benefit analyses.

5.6. Chapter summary

In this chapter, I have provided an overview of my theoretical framework and a description of my two-part research process. Using the social constructivist paradigm, I considered communities of practice and epistemic cultures to structure my research. The first part of my research process was conducting a literature review that inductively identified six key issues. During the second part of the research process, I conducted 30 semi-structured interviews with qualitative researchers, big social researchers, and data curators, using a questionnaire organized around the six key issues. The deductive and inductive coding process of the semi-structured interviews, conducted using ideas from grounded theory, resulted in nine broad parent themes and 104 subthemes that provide detail and granularity to the parent themes. In the next chapter, I provide an in-depth review of the results of the semi-structured interviews.

Chapter 6. Results

This chapter reviews the results of the data analysis described in Chapter 5. The results are structured by nine key themes. These themes include the six key issues that were identified in the literature review, as well as three more themes that emerged during my data analysis.

Original six themes

1. Context
2. Data quality and trustworthiness
3. Data comparability
4. Informed consent
5. Privacy and confidentiality
6. Intellectual property and data ownership

Additional themes

7. Domain differences
8. Strategies for responsible practice
9. Data curation issues

The 30 interviews with big social researchers, qualitative researchers, and data curators demonstrated that the original six themes identified in my literature review were the appropriate categories with which to group the interviews. Each group of participants had clear ideas about, and responses to, each of these six themes. Additionally, my post-interview deductive coding process revealed three more themes: domain differences, strategies for responsible practice, and data curation issues. These three themes proved to be analytically powerful lenses through which the participants viewed big social research and qualitative data reuse—how each community of practice understood their own disciplinary and methodological foundations and landscapes, the strategies that each community of practice used to support responsible practice, and each community of practice's experience with data curation.

I have organized my results by theme, with subsections that specifically discuss each community of practice—qualitative researchers, big social researchers, and data curators. These results are based on a sample of 30 total participants, consisting of 10 participants from each of the three communities of practice. An introductory paragraph in each section provides an overview of how each theme was addressed by each community of practice, including some instances of divergence or convergence of ideas across the three participant communities. In Chapter 7, I further synthesize the insights developed through the interview process.

Please note that while I provide numbers for how many interview participants addressed each subtheme, this number is not meant to suggest quantitative conclusions about the members of these communities of practice as a whole. I provide these numbers only to give a broad sense of how common it was for participants from each community of practice to discuss each subtheme. When numbers appear within a subsection, labeled "qualitative researchers," "big social researchers," or "data curators," those numbers represent that community of practice only. To improve the clarity and readability of quotes from the interviews, I have removed filler words and phrases such as "um," "you know," and "like" when they did not alter the meaning of the quote.

6.1. Context

The most common insight expressed by the participants regarding context for reused qualitative data and big social data was that documentation, description, and metadata can help preserve context (n=13; qualitative researchers (qr)=4, big social researchers (bsr)=2, data curators (dc)=7). However, many participants acknowledged that the practice of curating data and adding documentation also has two key drawbacks: (1) the process of creating thorough documentation, description, and metadata is time-consuming (n=7; qr=5, bsr=1, dc=1); and (2) the preservation and dissemination of contextual information is in tension with preserving participant privacy: the greater amount of detailed contextual documentation, description, and metadata are added to the data, the more likely the individual participants' data will be identifiable (n=10; qr=4, bsr=0, dc=6). The theme of privacy and confidentiality is discussed further in section 6.5.

Most participants had considered the idea of context, but different categories of participants (big social researchers, qualitative researchers, and data curators) spoke about different central concerns regarding context and different strategies for preserving the context of data.

6.1.1. Qualitative researchers

The qualitative researchers I interviewed tended to focus on how a researcher could influence or communicate the context of the data. Qualitative researchers discussed how their own methods, ideas, and values as researchers contributed to the context of the data (n=4). Some (n=2) suggested the strategy of collaboration with original researchers as a method for incorporating the original contextual expertise into a data reuse project. Others (n=3) suggested that a degree of misinterpretation may be inevitable and that context could never be fully communicated, but these researchers still considered the benefit of data reuse worth the risk of incomplete contextual information. As one qualitative researcher expressed it, “I have grown a bit of a thick skin in terms of my data and my publications being misinterpreted. Yeah, I do the best that I can [to provide contextual information] and then I just let it go” (QR02).

Qualitative researchers also discussed the tension between protecting privacy and preserving context (n=4). One qualitative researcher gave a detailed explanation of this tension in their interview:

> Do I say this was a group of people who are enrolled in an eating disorders program at [X University]? Well, now that could [allow the data to] be [re]identified. Someone could look at who's in the eating disorder program and maybe connect [a person’s] age to that. So I almost have to say it's the central [name of State] eating disorder group or something along those lines. That bothers me because if it's central [name of State], that means it could be urban. It could be [name of City] where I live right now, which is quite urban and Black and socio-economically divided. Or it could be central [State], rural, I have 500 cows, and I'm on a farm, you know. So it's really the context there. I have such an issue with that. And in telling enough context to be able

> to understand the situation and yet not give away the participants' identity or have any sense that there would be any identity accidentally misappropriated. So it is very hard. (QR08)

As illustrated in these examples, qualitative researchers conducted informal risk-benefit analyses throughout their research and data sharing processes, as part of their thought process about how to communicate context for shared data: What is the benefit of sharing data vs. the risk of future users misunderstanding the context of that data? What is the benefit of providing clear contextual information for the data vs. the risk of identifying individual participants? The practice of conducting this type of ad hoc risk-benefit analysis was mentioned by all three of the communities of practice I interviewed (n=17; qr=5, bsr=6, dc=6). The theme of risk-benefit analysis is discussed further in section 6.8.

6.1.2. Big social researchers

Big social researchers' discussion of context was often focused on the more technical aspects of context. They discussed the out-of-context effect of aggregated data—an effect arising from the fact that big social data are often merely small snippets of text or images that come from the broader context of a social media account as a whole, or a person's life as a whole. Some big social researchers (n=4) talked about how the big social data mining techniques remove the user interface as a contextual factor, leaving just text, image, and metadata. As one participant said, "If you only look at the text [of a tweet], you're stripping out a bunch of the context… The way that the API returns it to you, that's not how it's being seen in the wild" (BSR04).

Big social researchers also talked about structuring their research design and methods so as to support clearer context (n=3). For instance, one big social researcher discussed how their research used book-reading data from a social media platform in a way that was similar to how the platform itself used that data, saying "I think [the way we use this data in our research] is pretty faithful to the context of what's happening with the data in its original situation" (BSR03). Big social researchers (n=2) also talked about selecting data that had more inherent context, such as selecting Tweets that included a geographical location tag.

Unlike the qualitative researchers and data curators I interviewed, big social researchers did not discuss the tension between providing contextual information and protecting user privacy. See section 6.5 for further discussion of the theme of privacy and confidentiality.

Representativeness of the data was also a key topic for big social researchers (n=4). These researchers selected social media platforms that could provide the data they needed, but they were aware that the users of any single social media platform are not representative of the population as a whole. This concern about the representativeness of social media data was also discussed in relation to data quality and trustworthiness (see section 6.2).

6.1.3. Data curators

Data curators were most likely to talk about documentation, description, and metadata as a strategy for preserving context; 7 data curators discussed this topic, compared to 4 qualitative researchers and 2 big social researchers. Data curators identified context as a key to understanding archived data (n=2), and they also emphasized the importance of preserving related materials alongside archived data (n=4). One data curator suggested that web links within social media posts could provide context, but that "[web] links are a terrible type of data to publish. So we always do Perma.cc,[3] hoping that will be around longer" (DC09). This strong focus on the value of digital preservation, in addition to sharing and reuse, was unique to data curators.

Like qualitative researchers, data curators discussed the tension between providing contextual information and preserving privacy for human subjects (n=6). As one data curator phrased it,

> You're dealing with human subjects. You're concerned with potentially identifying them, and you have to follow certain guidelines. And in doing so, you remove a lot of the context that exists in those datasets to begin with. … And I have mixed feelings about that, because the scientific community has a lot to gain from having at least the fullest picture that they can take away from qualitative datasets. (DC10)

[3] https://perma.cc/

Data curators discussed this tension between context and privacy more than the other two communities of practice. Six data curators mentioned this theme, as opposed to four qualitative researchers and zero big social researchers.

6.2. Data quality and trustworthiness

Documentation, description, and metadata was the most commonly-discussed theme related to data quality and trustworthiness. All three communities of practice (n=18; qr=5, bsr=6, dc=7) discussed the care they took to fully describe any data quality issues so as to facilitate and support data reuse. Similarly, all three communities also indicated that they were more likely to find data trustworthy for their own reuse when quality issues were well-described in the datasets. All three groups (n=10; qr=1, bsr=4, dc=5) also touched on the idea of data completeness as an essential element of quality and trustworthiness, pointing out that high-quality datasets should include clear communication of which data were used in their analysis, which data were archived, and which data might be missing.

However, aside from the two themes I have just described, ideas about data quality and trustworthiness did not overlap between the three different communities of practice—qualitative researchers, big social researchers, and data curators. Rather, as I describe below, each community emphasized its own specialized considerations pertaining to data quality and trustworthiness.

6.2.1. Qualitative researchers

For qualitative researchers, the concept of using documentation, description, and metadata to communicate data quality usually referred to a discussion of quality in their manuscript, rather than a readme or other descriptive metadata that would be included alongside their published data (n=5). As one researcher explained, “I actually was able to write, ‘these are my methods, these are my interview guides. These are the steps that I took to enhance rigor’” (QR03). Another researcher emphasized their effort to expressly note in their manuscript whenever the data had been changed in any way, saying:

> I think I actually went into more detail in the paper that was linked to the [shared] data. And that's where I described a little bit more about how I went through and

> changed these transcripts. Basically, I used an online transcription service for recordings, [but] those have a bunch of random gibberish in them… And then, when I got into the [section of the paper in which I discussed] deidentification, I talked about the changes I made, trying to make it really clear: these are the kinds of things I changed, and this is how you know that I changed something. (QR07)

Beyond documenting their methods for data collection and data analysis, and documenting the completeness of their data, qualitative researchers' main concern regarding data quality and trustworthiness related to the inherent messiness of conducting research with human participants. Qualitative researchers discussed the difference between using transcripts or videos as data sources as opposed to talking in person with research participants (n=2); as one researcher explained, "Our video quality is okay, [but] it's not the greatest… We tried to think about doing some multimodal analysis, [but] it's just a little tricky with our video quality. There are things that you miss, right?... Facial expressions, smaller nonverbal cues" (QR06).

Qualitative researchers were also concerned about the degree of trust they could reasonably place in the original data creator when reusing data (n=2). As one qualitative researcher said about reusing archived qualitative data from previous eras, "There's a very well-documented history of racism in ethnography, and colonial foundations of ethnography" and "One presumes, one hopes, that there was an appropriate relationship there [between researcher and participant]" (QR04).

Lastly, qualitative researchers were aware of researcher bias—the ways in which the researchers themselves could affect the qualitative research process (n=3). As one interviewee explained, "We, as a [co-author] group, tend to value more highly the opinions of non-managers. I want to say: we have managers in our dataset, and they're lovely people. But [one] part of the impetus for [our] study is we're really sick of just seeing reports with managers saying 'The future of [the field],' [and] talking about labor without actually, like, doing labor, or caring about employees. So that is also, I guess, a more unconscious bias" (QR01).

6.2.2. Big social researchers

Big social researchers spoke about data quality and trustworthiness more than the other types of interviewees. Regarding documentation, description, and metadata, big social researchers generally focused on including code and calculations to document data quality. One researcher described their reasoning for documenting data quality in this way:

> In our doc we definitely have, 'this is where this calculated field comes from. This script comes from there.' If you want to poke and you want to change how we calculated those fields, you can do that, if you don't trust us to make those, or you want to do it a different way. So that was also something that was important for us. (BSR02)

As noted above in section 6.1., big social researchers again spoke about the representativeness of social media data, highlighting that using a non-representative dataset affected data quality (n=3). Big social researchers (n=6) also discussed spam and bots—how to filter out spam and bots, whether spam and bots affected the data quality, and when bots might be relevant to their research question. One researcher working with Wikipedia described bots that had a specific purpose on the platform: "[There are] these pro-social bots that are authorized by the community. And … some of them do a lot of routine maintenance work, find-and-replaces, cleaning stuff up" (BSR02). A few big social researchers used computational methods to filter out spam (n=3), whereas others were aware of spam but decided that their research didn't necessitate removing it (n=3). As one researcher explained, "I include [bots] as part of the dataset and see whether it could be an influential central entity in the social network. And in most cases, it doesn't become so popular in the network. But if a bot is identified as one of the central figures in the network, then I want to look at it more closely" (BSR10).

Other data quality and trust-related themes discussed by big social researchers included quality issues that arose with large-scale and automated collection (n=3)—issues such as cleaning up unicode or other programmatic quality issues, as well as problems with automated clustering or other methodological issues. Big social researchers also discussed combining datasets to support data quality (n=2); one researcher collected Reddit data using

a third-party app as well as through the Reddit API; another researcher compared “results generated from the Twitter data… with information collected from news articles. Because usually news articles are trustable. So we use information from news articles and government reports to validate the information we gathered from Twitter” (BSR09).

Other big social researchers (n=2) discussed how big social data are subject to loss over time—social media users can delete their accounts, links can become broken, and platforms can change. As one researcher described, “There's a paper that gathered a bunch of tweets, both related to specific events, and then just a broad sample of Twitter. And then five years later, they tried to re-access the same data, and they found—I think it was [only] about 75% of the tweets were still there. So in five years, they lost 25% of their data” (BSR06). Big social researchers were the only community to look to existing literature for guidance on data quality (n=2), reading similar papers to see how data quality issues were addressed.

6.2.3. Data curators

Data curators were less focused on documenting the quality of the data—which they viewed as outside of their purview. Instead, they focused on the quality of the documentation, description, and metadata itself. As one data curator phrased it, a full “description of the process, I think, should enhance trust for secondary users. [A description would ensure that secondary users] know what happened. Whether they agree that it was a good process or methodologically sound or whatever, then it's up to them. That’s, I think, who should judge quality. But the process description is fully there, and you can kind of follow it” (DC09). Another data curator concluded, “Our main impact on quality is actually the quality of the documentation and description, rather than the quality of the data” (DC02).

Data curators also discussed quality issues related to large-scale automated data collection (n=4). One data curator who collects tweets for archival purposes described data collection issues resulting from how the Twitter API changes over time:

> The API returning a retweet has only been possible since Twitter introduced the retweet button. And they have actually now introduced this quote tweet button. And they've changed the functionality of retweets. So that field from their API has

> changed as different versions of the API and different versions of Twitter software have been released. (DC08)

Another data curator reiterated the idea that a curator's responsibility regarding data quality does not extend to the content of the data: "It's [often that] a million rows have the same sort of data. So it's more just like, does this file load properly? Does it run through the related code properly? And are there any major issues in the metadata that I need to be concerned with?" (DC10).

Data curators were the only community of practice to discuss the idea of curator review as a strategy to support data quality and trustworthiness (n=4). One curator explained that "when a dataset is submitted to our institutional repository, at this institution, we check it pretty thoroughly for anything that might be missing, that may make it unusable or non reusable" (DC01). Another suggested, "I think our role is to be a somewhat neutral party" when reviewing data for publication (DC02). A third curator who works at a data repository described "a quality control process that we go through before any datasets get released. So for a qualitative study, a senior curator in the unit would review the dataset and the work that's been done, and then a supervisor would release it. So there are multiple eyes on it, in case anything gets missed" (DC05).

6.3. Data comparability

Participants from all three communities of practice were generally aligned on issues related to data comparability. Qualitative researchers, big social researchers, and data curators all discussed the challenges of interoperability, including data formats, metadata standards, language, encoding language, and other factors (n=10; qr=1, bsr=2, dc=7). Participants from all three communities of practice also emphasized the importance of being able to compare and combine data, noting that more data could lead to stronger research conclusions (n=10; qr=2, bsr=6, dc=2). All three communities of practice also discussed how documentation and metadata could support data comparability (n=8; qr=2, bsr=2, dc=4).

6.3.1. Qualitative researchers

Qualitative researchers had not considered data comparability as much as other groups. They discussed documentation as a strategy to promote comparing and combining datasets. One researcher explained, “We did publish our interview guide. And I think that actually goes a long way in facilitating interoperability, because people will be able to see the direct questions we asked and will be able to see whether or not the potential answers would be able to mesh with, for instance, other interview [data] or particular survey data” (QR01). This researcher also discussed interoperability (n=1), saying, “We wanted the format to be very similar. So all of our datasets have the same format. If you see something redacted, it all appears the same from a machine actionable standpoint. They're all very interoperable” (QR01).

Qualitative researchers believed that increased amounts of data could lead to better conclusions (n=2), and they saw their data as potentially complementary and combinable with quantitative data. For example, a researcher said, “This project was designed with the intent that it would complement the more structured data collection and analysis methods that the organization tends to use. So we know that the organization already has access to large sets of data that speak to the same issues” (QR02).

Two qualitative researchers also pointed out that the complexity of qualitative data could hinder comparability (n=2), giving examples relating to the inherent flexibility of semi-structured interviews. As one researcher explained, the use of a semi-structured format meant that "each interview within the same study can [potentially] be asked differently, and different prompts can happen. So unless you're doing a totally structured interview, which happens very rarely in my line of work, [comparability is difficult]” (QR08). The other said, “The [interview] guide is really just what it says—it's a guide. It's not a one-to-one question and answer. So that also can sometimes be a problem with interoperability in qualitative spaces” (QR01).

6.3.2. Big social researchers

Big social researchers emphasized how more data could lead to stronger conclusions (n=6). One big social researcher explained that the standard in their field was to use multiple data sources: “We have these three different sources of data. And that's partially because the recommender systems research community likes seeing results on multiple datasets” (BSR03). Another researcher described using a combined dataset to ensure that the Twitter accounts used in their research belonged to people in the United States: “They have public voter registration files… in the United States. So they match those to Twitter accounts. So what that does is it brings in the demographic information with the Twitter account, so you can start to ask questions like, what are real people doing on Twitter versus this weird mix of real people and bots and organizations and stuff like that” (BSR05).

Noting the benefit of more data, big social researchers also discussed the challenges of matching up different datasets (n=4). As one researcher told me, “Matching names is a difficult thing because of informalities and stuff like that, multiple people having the same name and same location” (BSR05). Another researcher further described the difficulties of matching up datasets: “You have to do something like a fuzzy text match. The good thing about this dataset was it was small, so I could manually inspect every single match to make sure that it's right. So I could check for false positive matches, but not for false negative matches. And so if I did not find a match, I didn't actually go and search for it manually” (BSR01).

Big social researchers were also concerned with interoperability (n=3). One researcher described a project that looked at Tweets in different languages, saying, “Even though [it was] the same platform, there are users of different communities who speak different languages. You… need a coder who understands those languages, so you need a team helping you. Or you would need to use technology such as Google Translate API or Bing Translate or any other platform, but either way, you will need help from humans or technology to assist” (BSR08). Another described a study of fake news on Twitter: “I have to search through all these Twitter data in my database. Then Snopes.com will have its own title for that fake news event. But if you use that title as is, … you will only collect tweets

that reference Snopes.com exactly as the title says. So I had to develop some strategy to use some synonyms of some certain keywords of these titles of the fake news" (BSR10).

6.3.3. Data curators

Among the three communities of practice, data curators were the most focused on interoperability, documentation and metadata, and the idea that combining data can lead to stronger conclusions. Regarding interoperability (n=7), data curators tended to consider file formats and metadata standards. As one curator said, "We always try and ask for nonproprietary file types, so plain text, CSV, that sort of thing. So that it's as interoperable as possible with as many different types of other data" (DC01). Another curator who worked at a repository described the use of standardized metadata formats: "To have standardized metadata we use a simplified DDI [Data Documentation Initiative] codebook. But we also have clean mappings to DataCite [metadata schema]. And especially with the most recent DataCite kernel updates, I think we can map almost any metadata field to DataCite's" (DC02). A third curator who was embedded in a research team walked me through their team's thought process when assessing interoperability: "Are our date formats the same? Is our blinding mechanism the same? Is our blinding good enough? Do we have confidence in our coding? Did we keep the data dictionary the same for the coding? Or has it changed over time? If it's changed over time, why was that? [We ask ourselves these questions] to help the coders or to help those who would interpret the data later during analysis" (DC06). Another participant told me about an initiative to support interoperability of different qualitative data analysis systems, saying: "We think about interoperability of qualitative projects that have been analyzed with software analysis packages like NVivo and Atlas. Because if somebody … doesn't deposit their raw materials … for one reason or another, but does deposit analysis output from some package? You know, that's good, that's better than nothing. But what if nobody else, or very few other people, have access to that same package?" (DC09).

Regarding documentation and metadata (n=5), curators discussed how documentation can help support comparability. One data curator spoke at length about synthesizing multiple qualitative datasets:

> Qualitative data aren't standardized the way survey questions are. And I don't think they should be, necessarily. So synthesis is harder, but can be done, right? You just have to be aware that questions were asked at a different time by different people in a different context. So you don't expect one-to-one, psychometric mapping of answers, but you can still understand similar trends and similar projects. So in that sense, coming back again to context, if you have enough context, if you understood how the data were generated, you can use them in a comparative reuse, or even a synthesis context. (DC02)

Another data curator described efforts to include documentation to support broader use of big social data: “If we have [the Twitter data] saved as JSON files, we're gonna have to do some training and maybe have a little Python script or something that can self-execute that you can run to take all of those that are in a directory and turn them into text documents that Joe Schmoe on his computer can read, without having to be a computer scientist” (DC04).

Two data curators discussed how more data can lead to better conclusions (n=2). One curator told me about a qualitative research project and a previous, larger survey of members of the military. The curator told me that the two datasets were natural complements, but that they were difficult to combine: “We have existing data, but except in a very limited number of cases, we don't have any way to link the [new] qualitative data to the [existing] survey, which gives us a lot more information about the people—everything from their rank and age and state of origin, what branch of the service they're in, all of that” (DC04). Another curator considered how large social media datasets can support longitudinal research: “I think it increases interoperability, and also the ease [with] which they could approach that research longitudinally, like pull that same data in a year, because it's more straightforward to do so” (DC10).

6.4. Informed consent

The issue of informed consent produced a wide range of themes, and the themes addressed by members of each community of practice were relatively distinct. All three communities discussed the role of the Institutional Review Board (IRB) as a review body that could

support ethical practice around consent (n=22; qr=7, bsr=9, dc=6). However, each community of practice viewed the role of the IRB differently, and these differences are explained below. Participants from all three communities of practice also touched on the idea that some big social data sources are considered more "public" by users, and therefore there is less need to be concerned about consent (n=3; qr=1, bsr=1, dc=1).

Other subthemes of consent were markedly divided. Data curators weighed in on most themes, but all of the remaining themes were discussed by either qualitative researchers or big social researchers, but not both. This result indicates that qualitative researchers and big social researchers have markedly different understandings of what informed consent means for their research, and different ideas about their responsibility toward research participants in terms of consent. These differences are discussed further below.

6.4.1. Qualitative researchers

While the majority of qualitative researchers (n=7) discussed IRBs as a resource to support ethical practices, qualitative researchers tended to be more skeptical of IRBs' ability to support ethical data sharing and reuse. As one qualitative researcher said, "I consulted with my IRB, and [their response was], 'What's the problem? [The data are] deidentified.' They don't *get* qualitative research… So I guess I didn't find the IRB very helpful in thinking through this question from an ethics perspective. They did let me know that I was off the hook in terms of an IRB [review]" (QR03). Another researcher whose consent procedures specifically addressed data sharing said, "We talked about a lot of different issues around [consent to data sharing] and decided that consent that would allow us to share data in the long run. And then we went to get it through the IRB. They had no—I was surprised. They didn't say anything" (QR05).

In another interview, a qualitative researcher worked through their mixed feelings when considering consent in a secondary analysis of quotes pulled from published research articles:

> My IRB said it's not human subjects research. … They gave me an exemption. And that, in some ways, made me feel like I at least had some… You know, I ran it by

> somebody else. So I did think about it. But I also thought, well, [the participants who were quoted in the research articles] went through an informed consent process [during the original research process], but I have no idea what that was like, other than people say, in their research articles, informed consent was obtained, right. So I didn't know what those informed consent forms look like. But I never felt like I needed to reach out and find out [about it]. I just felt like since they're publishing it, and it's available, it would be the… I don't know, it's so tricky. (QR08)

Beyond the approval process of an IRB, qualitative researchers also discussed how consent language and procedures affected data sharing and reuse (n=7). Qualitative researchers who were reusing data (either their own data or historical data) (n=4) found that the original data collection did not include explicit consent for data reuse, and therefore they had to make ethical decisions based on their understanding of the data. As one researcher who used historical qualitative data said, “I'm still reflecting on what is the most ethical way to engage with these data. … So for example, Indigenous societies for whom sacred or secret data are reported [in studies that did not use] what we would consider remotely appropriate consenting procedures today. And so the real ethical quandary is around the reporting of those data” (QR04). Another qualitative researcher wanted to publish their research data once the study was finished, but realized that their consent procedures hadn’t addressed data sharing: “We didn't get permission to put [the data] up [in a data repository]. So I guess we're just not gonna make our data available” (QR09).

Some qualitative researchers included specific consent to data sharing in their consent agreements, including some who included tiered options for consent to data sharing (n=2). One researcher described their tiered consent procedures: “We had different things they consented to. Like, we could use this data for just this research, project and analyses, or we could use it to share in other external presentations, or for other secondary purposes outside of this research project” (QR06).

Some researchers allowed participants to review and redact their own transcripts prior to publication (n=4). One researcher described the consent process for publishing qualitative interview transcripts as follows: “We said in the informed consent [agreement] that we're

going to send you a copy of the transcript that you will be able to redact. So that was very upfront with the participants. We said, 'we are really hoping that you will allow us to put this data in the QDR [Qualitative Data Repository], here's how it would look by going into the QDR, it wouldn't just be openly available, people would have to request it from us.' So we outlined the risk mitigation that we were doing by depositing in the QDR" (QR01).

Despite efforts to provide clear consent for data sharing, researchers voiced concerns about the difficulty of truly informed consent. Researchers suggested that no one can be sure how the data might be used in the future (n=5), and speculated that participants may not always understand the nuances of a consent form (n=4). Some qualitative researchers (n=3) were also concerned that openly addressing data sharing in the consent procedures could affect potential participants' willingness to participate in the research. As one researcher said, "Sometimes people say things in interviews that [aren't] particularly sensitive, but maybe they don't want to share [them] with the whole world" (QR09).

Qualitative researchers also mentioned that they felt there was a scarcity of guidance and ethics rules to help them navigate consent for data sharing and reuse (n=3). Many talked about developing their own personal strategies and goals for responsible practice; this idea is discussed further in section 6.8.

6.4.2. Big social researchers

Big social researchers generally looked to IRBs to provide an ethical stamp of approval for their research. Only one (n=1) big social researcher described a more in-depth interaction with their IRB; their study involved suppressing users' "reputation score" in an online debate community without users' knowledge. As the researcher told me,

> IRB specifically asked me a lot of questions about consent. So they were interested in firstly, are the people on the platform going to know that you're hiding their reputation? And for me, it would be bad if they knew, because that would change their behavior. So I didn't want to explicitly tell them. So I had to justify that … not having informed consent while doing the experiment was not causing a lot of harm. (BSR01)

The remaining big social researchers who mentioned an IRB (n=8) told me either that their project was given exempt status by their IRB, or that they did not submit the project to an IRB at all, since they did not consider their project to be human subjects research. As one interviewee explained, "The type of data that we get are publicly available data. So somebody voluntarily consented to post [that] information online to let the world see it. And so we … do not consider that these are studies that require informed consent, because technically, there are no participants" (BSR08). Some researchers (n=4) did not feel that informed consent was necessary for big social data because most social media terms of service include a broad consent agreement that users must agree to in order to use the service. As one researcher told me, using Twitter data without express consent from users "feels a little icky. But in terms of what actual regulations are there, we were leaning on … Twitter's Terms of Service and how they govern the use of these developer accounts that you have to [register for] to access this data. That was what we kept going back to say: 'Okay. According to these rules, it is okay for me to publish this data'" (BSR06).

Others spoke about their efforts to design their research responsibly, even without explicit consent from the people who are reflected in big social data. A few of these researchers (n=3) looked to ethics education and ethics-related literature as a guide. As one big social researcher told me, "I had read the AoIR [Association of Internet Researchers] guidelines. And so we had that as a common reference point. We knew that the IRB…considered it public data, they didn't care. We knew it was on our shoulders to take care of all of this" (BSR04).

Big social researchers described a variety of strategies and considerations regarding consent. Several researchers described taking care with direct quotes (n=4)—either altering quotes so as not to publish users' words verbatim and/or removing usernames. As one researcher researching on Pinterest told me, "[Users] do have certain expectations, or, it could be a lot more unconscious than that… if you ask them to stop and think about it, like, 'Hey, would you like to see this [Pinterest post] published in a journal?' then they would think, 'Yeah, I should give my permission for that to occur.' So I don't think it's right to publish usernames,

or if it was something else really identifying, like a picture of a person, I would definitely have second thoughts about that" (BSR07).

Big social researchers also considered the public or private nature of information posted online (n=3), considering consent to be less of an issue for data that could be considered more public or users who were public figures. One researcher described how they considered hashtags to create a more public online space. As he said to me, "I will say which hashtags I'm using—I'll identify the hashtags, but I'll be careful not to identify the users" (BSR04). Researchers also used the potential harm to users as a criterion for assessing whether consent was an issue in their research (n=2).

6.4.3. Data curators

Data curators discussed IRBs to support consent (n=6). Some viewed IRB documentation as a stamp of approval, or a way to encourage transparency for shared data—the idea that "showing the IRB approval does sort of guarantee that the people who are using it have certain ethical structures that they're following" (DC07). As another curator said, "A rough idea at our institution is, if we're going to house human subjects data, regardless of whether or not it's been [deidentified], we need an IRB number to go with it. So we're discussing whether or not that's going to become a permanent part of our [repository] metadata" (DC01).

However, other data curators considered the IRB's role to be more nuanced. A curator who works at a national data repository explained the challenges of dealing with IRBs from different institutions:

> IRBs are only of limited help here, because a lot of how your IRBs think once data are deidentified, they're no longer human participant data. And so they kind of wave their hand. Not not in a unified way, right. In the U.S. it's kind of 50/50; some IRBs say you can't publish data and some IRBs say no, that's fine. And we get into this weird situation where… the IRB says, sure, you can share that [data]. But we [the curators] don't really think you should. (DC02)

Another curator described the evolving ideas about big social research among curators and IRBs, saying “I'm on a professional forum that IRB personnel are [also] part of, and in [some of] the discussions that I've seen, IRB personnel really want the researchers to identify themselves. Not like, you know, post-fact scraping of stuff. To identify themselves to say, this is what I'm doing… basically to do some version of informed consent online” (DC09).

Like qualitative researchers, data curators also discussed how consent language and consent procedures affect informed consent (n=8). Curators told me about qualitative researchers approaching repositories to publish data without having clearly indicated in their research consent form that data would be shared. These cases were difficult for data curators to navigate; curators needed to weigh the risks to participants against the benefits of data sharing. As one curator said, “In some cases it's so clear cut, like it says very explicitly, me or my research team are the only ones we're ever going to see these data, identified or deidentified. And in those cases, [the repository] really just can't process the data. We then offer various creative suggestions of providing some transparency. [For example,] the code book part, and then several illustrations. So, unfortunately, we've published many, many projects like that” (DC09). Curators generally suggested that if the consent language that participants received didn’t specifically address data sharing, decisions could be made on a case-by-case basis about whether the data was still shareable; these decisions often depended on the sensitivity of the data (n=3) and whether the data were completely deidentified (see 5. Privacy and confidentiality). Data curators also suggested that research participants could be contacted to reconsent to data sharing, although data curators acknowledged that this happens quite rarely; it can be difficult to reach participants, especially if a substantial period of time has elapsed since the initial study (n=2).

Like qualitative researchers and big social researchers, data curators spoke about the conflict between “publicly available” data and participants’ expectations of privacy (n=4). Data curators had also considered the idea of archiving big social data, not just as research data, but as archival materials to support the historical record (n=3), and the concern that it is impossible to know how data might be used in the future (n=3)—an idea that calls into question whether consent to data sharing can ever truly be *informed.*

6.5. Privacy and confidentiality

Among the key themes identified in the interviews, the issue of privacy and confidentiality had the most consistency between the three communities of practice. Qualitative researchers, big social researchers, and data curators all had similar understandings of how issues of privacy and confidentiality factor into research, and many of the subthemes in this category were discussed by all three types of participants. All three communities of practice discussed considerations pertaining to data deidentification (n=18; qr=8, bsr=5, dc=5), data sensitivity and vulnerable populations (n=11; qr=4, bsr=3, dc=4), using restricted access to support privacy (n=11; qr=3, bsr=1, dc=7), participant/user expectations of privacy (n=10; qr=1, bsr=5, dc=4), consideration of potential harms (n=10; qr=2, bsr=2, dc=6), how research design can support privacy (n=8; qr=1, bsr=4, dc=3), and data security concerns (n=6; qr=2, bsr=2, dc=2).

6.5.1. Qualitative researchers

Qualitative researchers generally had well-established strategies for protecting the privacy and confidentiality of research participants, and these strategies did not change for data sharing and data reuse. Qualitative researchers discussed deidentification as a privacy-protection strategy (n=8), and noted the challenges of deidentification. A qualitative researcher who wanted to share their data openly said, “Because we wanted to put no restrictions on it, [the curators at the data repository] went through line by line for each transcript, and pointed out things that could be potentially reidentifying” (QR01). Another qualitative researcher described the time-consuming nature of deidentification of qualitative data:

> We actually had a three-part process for reviewing the transcripts. So we had a person who went through the entire transcript to remove names and to flag issues that we might need to remove, either because they were identified in context or because there was something about them that we felt was sensitive enough that the participant probably didn't really want it in there. Then the second person would go through that same transcript, double checking to make sure that all names were removed, and try to resolve the issues that had been flagged by the first researcher. And then it came to me. And at that point… I went through all of the issues that had

> been flagged and made determinations on how we were going to handle them. (QR02)

Qualitative researchers also discussed restricted access (n=3). For example, one researcher "required that anybody using my data [had to] show us some sort of training, like CITI training, some sort of IRB ethics training, and if they had that, then that would be okay. But I wanted that to be a prereq[uisite]" (QR03). They also discussed implementing data security measures for identifiable data (n=2). As one qualitative researcher described, "Three people had access to the raw data. It was me, the lead investigator and their assistant who… checked my transcribing. So … we had [an] Excel spreadsheet that was password protected [identifying] site one and site two [with] the specific city and the hospital [where data collection had occurred]" (QR10).

Qualitative researchers also considered potential harms to participants that could result from data sharing (n=2), especially for sensitive data or data from vulnerable populations (n=4). As one researcher said, "I do believe in open data. But I think that there are a lot of considerations about understanding the data and placing the data in context that I think are very important when you're looking at any kind of sensitive data" (QR05).

6.5.2. Big social researchers

Despite the fact that big social researchers generally did not consider informed consent necessary for their research (see section 6.4.2. for more detail), they showed a high level of concern about protecting user privacy. One strategy that big social researchers described for protecting privacy was deidentification (n=5). As a big social researcher told me,

> I've come up with a workflow where I'm very careful to not include identifiable information. And by that I don't just mean user names, but I try not to directly quote tweets, and if I do, then I have a darn good reason for doing so. And I make it so that I'm studying a phenomenon, but the unwitting participants in that phenomenon, I do my absolute best to make sure that my work cannot be traced back to them in any way. I feel that's really, really, really important. (BSR04)

Another big social researcher described their strategy for deidentifying tweets that would be included in their paper, saying, “We didn't report actually direct quotes, we altered the text. [To do that,] we mash together similar tweets, so that, hopefully, they shouldn't be identifiable. Like, you shouldn't be able to reverse look them up or something like that” (BSR05).

Big social researchers also considered participant expectations (n=5). One researcher who uses Wikipedia data in their research told me, “There is actually a page on Wikipedia of people who have opted out of … being in those lists of the most active contributors. So we can also take a look at that. And whenever I do a peer reviewed article that's Wikipedia research, like I'll always check that list” (BSR02). Another researcher described privacy considerations as a key tension in big social research:

> There are tensions between what I want as a researcher, and what I would want as someone being researched, and I tried really hard to iron out some of those tensions. And I've told you already, I try not to identify people. But at the same time, there's no getting around the fact that there's this concept of surveillance that I'm really uncomfortable with. And yet my research depends on related concepts, or arguably the same concepts in order to function. And if there were the kinds of privacy protections out there that I might like, I might not be able to do the research. (BSR04)

Other big social researchers described efforts to design their research from the beginning in a way that supports user privacy (n=4). One researcher described selecting a research topic “that is completely derivable from public data and does not involve any sensitive personal attributes. So we could catalyze this kind of research without creating new privacy or discrimination problems through making an archival dataset available” (BSR03). Another researcher said, “My focus will be more on more public entities like institutions, federal entities, public libraries, and FEMA. And maybe some [individual users’] tweets will be contributing to my topic modeling study, but I try not to talk about individual tweets, exposing their private information” (BSR10).

Big social researchers also considered the sensitivity of data (n=3). As one researcher explained, “I tend to be cautious, maybe overly cautious about this. With the populations that I study… I have looked at students’ tweets, I've looked at teachers’ tweets, I've looked at politically and religiously sensitive populations. I don't think I've ever felt comfortable [sharing the tweets I’ve collected]” (BSR04). Another researcher described talking with their colleagues about social media data for a study of a hashtag on Twitter relating to sexual violence: “We don't want people to be able to easily identify survivors of sexual violence…. We had several conversations about that among ourselves, trying to figure out if we could share the data responsibly” (BSR05).

6.5.3. Data curators

Data curators were especially concerned with repository and curatorial support as it relates to privacy. Curators discussed strategies for sharing data with restricted access (n=7), and discussed using different levels of care depending on the sensitivity of the data (n=4), including being more stringent about data security (n=2). One data curator described assessing datasets to determine what privacy protections should be implemented: “What types of sensitive information is there? Does the study involve minors? Does the study involve other vulnerable populations? Can this data be linked to other people? Is there information on other people … that weren't the respondent? … how harmful would it be to the participants if this data were to be breached?” (DC05).

Similar to the process DC05 describes in their quote above, several data curators considered the potential harms of identifiable data, and used that criterion to make decisions about privacy-related data sharing strategies (n=6). One data curator described their decision not to share a dataset of GPS data derived from fitness trackers, “Considering the danger, even if the data is anonymized. I mean, just think about putting a map in a paper somewhere with ‘Hey, look, here's a point where 25 to 30 women in the dark of night run at the same time” (DC01). Another curator described conducting data reviews to identify any risk of participant identification: “The study actually was initially set to be a public release, so that pretty much anybody … could download it. But through my review and communication with my supervisor or the project manager, and then with a PI, we decided no, this is just too

sensitive. You're able to reidentify participants too easily just as it is, to be able to [make the dataset] public. So it was changed to a restricted access release" (DC03).

Curators described providing deidentification help to researchers (n=5), but were also aware of the challenges of deidentifying qualitative data. One curator described how qualitative data, even when thoroughly deidentified, is still identifiable by the research participants themselves.

> I think there is a particular perhaps unexplored issue with qualitative data—and this maybe similarly applies to to social media data—as opposed to [quantitative] data. Participants would always be able to recognize themselves in deidentified [qualitative] data, right? If I see a survey, and it's been deidentified, I cannot find my role. If I see 100 deidentified transcripts, it takes me 20 seconds to to recognize mine, which means participants know that their data is in there if they ever were to access it. (DC02)

Data curators also discussed participant expectations around privacy and confidentiality (n=4), for both shared qualitative data and big social data. One data curator described respondent expectations for shared qualitative data: "[Respondents are] agreeing to be anonymized, but they're also providing all of this extra detail. What were their expectations? It's sometimes hard to [know], especially if they're not coming from a research-oriented background. Are our expectations the same?" (DC03). Another data curator focused on user expectations regarding big social data: "There's an interesting thing that occurs when [deidentified] user data that people have consented to being collected, is made public. … [It] exposes the fact that the data is being collected in the first place, if that makes sense. That will often elicit this fearful or shocked response from the general community when they're like, wait, we didn't know that you were doing that" (DC10). Another data curator described the privacy implications of a dataset of Tweets that used the #MeToo hashtag. This hashtag gained traction on social media in 2017 and was associated with a movement calling attention to sexual assault and harassment (Bogen et al., 2021; Walsh, 2020).

> People who are using the #MeToo hashtag, some number of folks who use that hashtag, were really putting themselves at risk of backlash or harm by using that hashtag. And yes, they did use a public hashtag on a public forum. So none of these

> are private tweets with the hashtag, they're all public tweets with the hashtag. But a user who's participating in a large international discussion about what's appropriate in the workplace and what's appropriate for how we treat other people and their body autonomy has, I would expect different expectations about who will access that data and in what ways than a public figure making a statement on a public forum. (DC08)

6.6. Intellectual property and data ownership

Compared to other key issues identified in this dissertation, issues relating to intellectual property (IP) and data ownership were less clearly understood by the participants. A common idea discussed in interviews was the participants' lack of clarity about IP rights and data ownership (n=5; qr=1, bsr=2, dc=2). Members of all three communities of practice also touched on the idea of purchasing or using commercially-available data as a strategy for resolving IP and data ownership concerns (n=8; qr=1, bsr=3, dc=4). Participants also discussed data licensing (n=6; qr=3, bsr=1, dc=2) and data citation (n=5; qr=3, bsr=1, dc=1). Some also suggested reaching out to participants and organizations involved in the original research to discuss data reuse, although this strategy was only mentioned by one member of each community of practice (n=3; qr=1, bsr=1, dc=1).

6.6.1. Qualitative researchers

Several qualitative researchers discussed data sovereignty and ownership when considering sharing or reusing qualitative data (n=5). One researcher told me, "I think [my institution] tends to look the other way when [data] isn't patentable" (QR05). Another researcher said, regarding "the intellectual property of the people who are in the studies, … I confess that I had never thought about it that way until I started to learn about the Indigenous data sovereignty literature. And that was this total worldview shift, and it got me thinking about data in a very different way" (QR04).

Data citation (n=3) was mentioned by qualitative researchers as a strategy to protect intellectual property rights and acknowledge data ownership. For example, one researcher said, "We have, in the readme document, a statement that says how you should cite this

work" (QR06). Qualitative researchers were also aware of data licensing (n=3) as a strategy for informing others how shared qualitative data can be used. One researcher who had shared data in a data repository described sharing some of the data openly, and some with restricted access; they said, "When we published the open data, I believe, [it was licensed] CC BY [Creative Commons Attribution license]. The closed data is subject to [the repository's] specific terms of access, plus whatever we've added on to it. But … the actual IP remains with the [data creators]" (QR01). However, another qualitative researcher believed that data was not licensable, saying, "We cannot license our reports or the data or anything; it's not allowed" (QR02). Although only one qualitative researcher specifically mentioned confusion about IP rights, these conflicting quotes from participants illustrate the participants' limited understanding of IP laws, especially how they apply to data sharing and reuse.

6.6.2. Big social researchers

Big social researchers were most concerned with IP as it relates to using data derived from commercial entities. Big social researchers discussed the terms of service imposed by social media platforms and data providers (n=8)—usually trying to follow these terms of service, but sometimes making calculated decisions about when to bend them. Describing following the terms of service, a big social researcher said, "[In] the data management plan, I specify that I'm going to share [what] data I can, but note that some data is not going to be shareable either due to upstream restrictions—several of the datasets I'm linking, I'm not allowed to redistribute. Almost anyone can go get the copy themselves, but I can't provide it" (BSR03). Another researcher had gone against Twitter's terms of service to conduct web scraping for a subset of data, telling me, "The terms of service aren't ethical rules. They're just a set of guidelines set by a corporate company to protect themselves" (BSR05). Another researcher described the difficulty of adhering to terms of service that change regularly: "We were using [the Instagram] API before they changed the user agreement. I think, after a certain—and I forgot at what time, Instagram changed the agreement, and severely limited the volume of information that a researcher can download… And so the published research that involves Instagram actually cannot be repeated in the future" (BSR08).

Big social researchers also discussed purchasing or using commercially available data (n=3). One researcher described their use of datasets that had already been collected and posted online by other researchers: “From an intellectual property liability perspective, the people who scraped and initially produced the data would be on the hook. That's one reason I'm not redistributing the data… the datasets are very well known and are still available…. It's one of the reasons I've been hesitant to do a bunch of scraping myself—is just to avoid that set of issues” (BSR03).

Like qualitative researchers, big social researchers lacked a clear understanding of IP laws and were hesitant to speak in detail about them. One participant said, “Because I'm not a legal scholar, I don't know if Fair Use applies to the concept of violating the terms of service agreement” (BSR04).

6.6.3. Data curators

Data curators had similar concerns and strategies for dealing with intellectual property as other participants. They discussed social media terms of service—both following them and bending them (n=5). Data curators also talked about purchasing or using commercially available data (n=4). For example, one data curator said, “Just yesterday, we had an inquiry: ‘I want to do a sentiment analysis on 2000 Wall Street Journal articles from the Factiva database. I see they have an API, can you help me?’ Well, no, I can't, because we're not legally allowed to do that with our agreement. But if you have a few thousand dollars and would like to share it with them, I'm sure they'll help you” (DC04). Another data curator described handling copyrighted material in a data deposit: “The data producer included a copyrighted instrument, … but they've included that data within the dataset and within their full questionnaire. And so that was just me going back to the PI (Principal Investigator) and being like, ‘Hey, was this supposed to be released? … Did you have permission to to include this with your deposited data?’” (DC03). One data curator described their repository’s data enclave strategy for protecting privacy and IP rights for big social data: “We'd like folks to bring the analysis to the data. And then we'll review the analytical output for disclosure risk, just like we do with qualitative research studies. And so instead of reviewing all of the data on ingest, we review all of the results on download” (DC08).

Other data curators talked about data sovereignty and ownership (n=2). One data curator said, “I really like that idea of community-driven data governance… you can't do that, in the case of [qualitative datasets that are controlled by private companies]. Or really, either in the case of big data, because it's so disconnected already. But when you're working with new qualitative data, when you're talking to people to try and find those ways to let people have a say. It’s not only informed consent, but later, [asking,] ‘Do you think this represents you?’” (DC04). Another data curator touched on data ownership for academic researchers, saying, “The data technically always belong to the institution, even if researchers don't realize that” (DC09).

Data curators also discussed repository terms of use (n=2), and data licensing (n=2) as strategies to support intellectual property rights. For example, one data curator described the terms of use at the repository where they work: “We have a standard download agreement … it's essentially education and teaching, only non-commercial use, no brand production, no attempts to reidentify participants. Those are the key points” (DC02).

Like qualitative researchers and big social researchers, data curators also had a lack of clarity about IP laws (n=2). One data curator worked through ideas regarding IP: “I know that the legal situation is maybe a little gray. I think it's clearer in the US… and I think UK Data [Service] is more worried about this, I think they have actually built in copyright transfer, or some license, into their some of their consent forms. I would worry that that's a deterrent and also potentially unethical. And unclear what that even means for an interview. So I'd worry about writing too much legalese in there” (DC02).

6.7. Domain differences

The theme of domain differences emerged during my deductive coding process. The term “domain” is a term used by Wenger et al. (2002), and is an element of their idea of communities of practice. “Domain” describes the combination of interests and disciplines that are present within a community of practice. All three communities of practice in this study mentioned key differences in their behaviors, attitudes, and practices. Qualitative

researchers, big social researchers, and data curators all referred to data sharing values and norms (n=12; qr=7, bsr=2, dc=3), research practices and standards (n=9; qr=1, bsr=4, dc=4), and skills, training, and background (n=8; qr=4, bsr=2, dc=2) that were specific to their respective communities. Both qualitative researchers and big social researchers talked about collaborating together to support scaled-up, responsible research (n=4; qr=1, bsr=3, dc=0).

6.7.1. Qualitative researchers

Regarding community-specific research practices and standards, one qualitative researcher explained to me their guiding philosophy of qualitative research: “When you are sitting down with someone, and they're telling you a story, they’re giving you this gift of their knowledge and their experience. And I think qualitative researchers as a group have been really thoughtful about acknowledging the value of that ideology of respecting respondents, and wanting to do right by them” (QR03).

Qualitative researchers generally assumed that anyone reusing qualitative data would be trained as a qualitative researcher, with the accompanying skills and background (n=4). For example, one qualitative researcher explained why they didn’t include an explanation about sampling bias alongside their dataset: “I guess that is disciplinary bias, right? I assume that if you want to use this kind of data, you've had a basic methods class in anthropology or sociology, [and] you already know what some of the weaknesses of this are” (QR02). Another researcher explained to me that qualitative researchers themselves are a key part of the research data, saying “a core part of qualitative research is the idea of researcher as instrument” (QR03).

Qualitative researchers described a general reticence to share data among their fellow qualitative researchers. However, many I talked to were interested in trying to share (n=7). One researcher told me, “I'm an editor of [an academic journal]. And I find people not even wanting to provide their codebook because they're like, ‘That's not the essence of qualitative research.’ And I'm like, well, then how can we ever analyze or determine what kind of paper you're producing if you don't even want to give us the codebook? So I think there's gonna be a lot of hesitancy for people to also give up the whole interview, [even] if

it's deidentified" (QR08). Another researcher described their own concerns about sharing data: "I guess you just have to hope that people aren't going to A) misinterpreted it, or B) rip it to shreds for something… it does make you vulnerable when you put your data out there" (QR09). Conversely, one researcher argued in favor of sharing data: "Many of [the participants in my study] said, 'I want to help other people. I want people to learn from my experience. I want to share this.' And so I do have that in mind… when I said to you, 'Why shouldn't other people do more with this, as long as they're going to be responsible and respectful?' I feel like that's making more use of it" (QR03).

Only one qualitative researcher talked about collaboration with big social researchers (n=1), but they reported a broader adoption of collaborative practices: "A lot of the people who I know are working [with social media] are computer scientists. So for us, as qualitative researchers, we are always looking at what computer scientists are doing, and trying to figure out how we can use these innovations" (QR04).

6.7.2. Big social researchers

The big social researchers I interviewed discussed different practices and standards of different communities of practice (n=4). For instance, some researchers described a potential conflict between the common practice in their field and their own sense of responsibility, but they ultimately chose to stay aligned with other researchers in their area. As one researcher said, "There is a contextual integrity thing here. When the user submitted the review to [the social media platform], using it in my research wasn't their intention. But we are working entirely with public records. This is standard practice for recommender systems research. There's good arguments that perhaps it shouldn't be, but it is standard practice" (BSR03). Another researcher who was trained as a journalist said, "I think a lot of it was the training that I received in the [journalism] program. We talked about [big social data as content, rather than human subjects data] in our quantitative methods class. But I have some qualms about just saying, oh, we're studying content, we're not studying people" (BSR07).

However, another big social researcher described how a previous experience working with social scientists on human subjects research informed their current Twitter research:

> This is really sensitive, fully identifiable data. You have a name, you have an address, you have when their power went out, when the power came back on. … If possible, if there is something that could in any way be considered sensitive, it makes sense to deidentify. So having done work with PII [personally identifiable information] led to this idea that maybe this isn't PII by the letter of the law, but it is PII—sensitive adjacent. And so it felt like the right thing to do, [even if it was] not necessarily governed by something. (BSR06)

Other big social researchers also discussed the idea of collaborating with social scientists to support responsible practice (n=3). As one big social researcher who was trained as an engineer told me, "We interacted with and used a lot of expertise from some people in [the] communication [discipline] to try to have a better sense of it. As an engineer, that's something that would totally get washed away. And so we really wanted to make sure [our research] was grounded in communication or sociological theory" (BSR06). Another researcher described the benefits of multidisciplinary research: "Since my … graduate student years, … all my projects were multidisciplinary. So I had many chances to learn from sociologists and environmental scientists, geologists, and people from many different fields. So over time, I developed my current strategy and a set of tools to look at this social media data" (BSR10).

Big social researchers also discussed how different communities of practice have different skills, training, and backgrounds (n=2). As one researcher said, "It was a tough collaborative effort to try to find people who could be at this intersection. To be programmatic enough to pull 150 million tweets from Twitter, the Venn diagram of the people who can do that, and the people who have firm social scientist training and understand what this data means, is vanishingly small. And so it was a lot of collaboration and a lot of discussion to try to create a team that could balance both of those" (BSR06). A public health big social researcher described residing in a liminal space between computer science and social science:

> The type of research that I have done is … not the type of thing … where you have supercomputers doing deep learning and discover something that we can't really

> consider, but a computer algorithm can generate. I mean, I'm not a computer scientist. But at the same time, I'm not doing the type of traditional qualitative research where people are wanting focus groups, or one-on-one interviews, providing a lot of context to the specific tasks or specific documents or specific social media posts that they generate. (BSR08)

Another subtheme related to community-specific data sharing practices and norms (n=2). One researcher described this idea in detail:

> I wonder how much different disciplinary norms [affect data sharing]. I think the Open Science movement is largely fueled by the hard sciences. And when it comes to the social sciences, … you've got a chunk of social scientists who want to be like hard scientists, and so take a lot of cues from them. And then a whole spectrum going all the way to social scientists who are informed more by the humanities. And the set of values and priorities is pretty different. And I think this is especially true in education, where you have researchers who are informed by sociology, but also researchers who are informed by psychology and taking cues from the hard sciences. And so sometimes you butt up against each other about the very assumptions of what research is and what values [you have]. … And I think about that a lot when I'm trying to balance these open science ideals with other ideals. (BSR04)

6.7.3. Data curators

Data curators were able to speak about the differences between qualitative researchers and big social researchers from an outside perspective. Among the 10 data curators whom I interviewed, there was a variety of experience working with both big social data and qualitative data.

Regarding differences in research practices and standards (n=4), one data curator suggested that “the people we work with [who have] big data are usually data scientists, computer scientists, engineers, people who think in big boxes and mechanisms and are taught less to be attuned to the human consequences” (DC04). Another data curator described a similar perception of the difference between big social researchers and qualitative researchers: “For

me, the biggest difference is the relationship between researcher and participant. … How qualitative researchers talk about their participants and their relationship to participants and what that means for data sharing both on an ethical and protection level, but also on an epistemological level" (DC02).

That same data curator continued on to connect the differences between these communities of practice to differences in data sharing norms (n=2), saying, "I think that's so essential for how qualitative researchers think about sharing the data and why many of them are reluctant to share the data. Whereas with social media researchers, I think it's often us in repositories, and our lawyers, who have to put on the brakes, because they're like, oh, let's just take all of OkCupid and just put it out on GitHub" (DC02).

6.8. Strategies for responsible practice

Another theme that emerged during my deductive coding process was identifying the different strategies that participants used to support responsible practice. As mentioned above, all three communities of practice talked about the idea of conducting informal risk-benefit analyses throughout the data collection and data sharing process (n=17; qr=5, bsr=6, dc=6). All three types of participants also told me that they relied on discussions with colleagues and collaborators to work through ideas and decide how to support ethical, legal, and epistemologically sound paths forward (n=9; qr=4, bsr=4, dc=1).

6.8.1. Qualitative researchers

Qualitative researchers were aware of trying to balance the benefit of their research with any potential harms to participants. As one researcher described it, "The gift that we've been given is [participants'] time and their sharing of knowledge. And so the same instinct that makes us protective—we don't want people to be harmed—also makes us want to do the most with the data and make it the most helpful. And so sometimes that's where you end up. Being in a place where there's a conflict between those two things" (QR03). With few formal guidelines about responsible practices for qualitative data sharing, qualitative researchers looked to colleagues and collaborators to discuss ethical, legal, and epistemological concerns. The informal nature of these discussions is captured by a quote

from one researcher who said, “I did hit up my friend who has a PhD in history and used to be the qualitative specialist at [a major university]. And I said, ‘Would it totally invalidate our study if we let our participants redact their own transcripts?’ And she's like, ‘No.’ So I just took her word for it” (QR01). Another researcher described conversations about transcript deidentification, saying, “We kind of came up with our own protocol. We looked all over, there's really no protocol for deidentification” (QR03). Another strategy to support responsible qualitative data reuse was described to me by one researcher, who said, “You need to confine the conclusions. You draw [conclusions] very, very strictly and carefully to what the data can and can't tell you” (QR04).

Qualitative researchers were also most likely to discuss how the power dynamics of research could affect responsible data sharing and reuse (n=3). One researcher described specific challenges of their research: “When you show up as a researcher with the organization, and one of the two highest officials in the organization is saying that they endorsed the research, first of all, you have to be very careful [to ensure] that people are [actually] volunteering. And second of all, it's possible that there is an assumption that the data are only going to be used by the organization itself. So we tried to be very careful, both in the consent process and in the way that we framed the access criteria, to make sure that people would use it appropriately” (QR02). Another researcher described how power dynamics within the research team influenced decision-making: “At the time I was a second-year PhD student where I was just like, ‘well, you're the experienced person. That's the way you've done it before. All right.’ So I deferred to the senior person on the team” (QR10).

6.8.2. Big social researchers

Like qualitative researchers, big social researchers weighed a variety of risks and benefits as they conducted their research (n=6). One researcher discussed replicability versus privacy: “If I don't release the data, it will be private. But then no one can replicate my results. It's going to be really hard because you need to collect all this data again. So that's the trade-off” (BSR01). Another researcher weighed the idea of informed consent against the potential risks to social media users: “We're trying to be careful that we're not exposing users to new risks… but we [didn’t get explicit] informed consent from the users whose data

we're using" (BSR03). Another researcher talked about weighing the risks and benefits of breaking social media terms of service: "I have been involved in projects where we have knowingly violated the terms of service and we have judged the benefit of doing so to outweigh the ethical fraughtness of that. … We've had a conversation about it, we've decided that it was worth it at the end of the day, and we went with it" (BSR04). A quote from BSR05 sums up the process that many big social researchers used: "We try to do it as a balance. Do we think this research is important enough? And … if we think it is important enough, what safeguards can we put in place to make sure that this person isn't going to face harm from being in the dataset?" (BSR05).

Most of the big social researchers I interviewed described having conversations with their colleagues and collaborators to work through ethical, legal, and epistemological issues (n=8). For example, one participant described the benefit of discussions with collaborators whose values were not aligned with their own:

> I have co-authors who are advocates of open science and the idea that you share your data with everybody. And we've gotten together to try and figure out which of these two research virtues—the openness versus the ethics—which do we value? … It's been really interesting to have those conversations together, and to hear from someone I respect [about] the importance of sharing our data as much as we can. But at the same time feeling strongly that sharing it globally, instead of on a more limited basis, is that the way to go? (BSR04)

A few big social researchers also looked to ethical guidelines, including the Association of Internet Researchers Ethical Guidelines and the Text Retrieval Conference's Fair Ranking Track (n=3). Big social researchers were also more likely than other groups to consider appropriately tailoring their research questions and research scope to support ethical, legal, and epistemologically sound practice (n=4). For example, one researcher told me, "We tried to go with [a research question] that is completely derivable from public data and does not involve any sensitive personal attributes. So we could catalyze this kind of research without creating new privacy or discrimination problems through making an archival dataset available" (BSR04).

One big social researcher who was studying a religious fringe group on Twitter discussed the power dynamics inherent in the research:

> So this is a misogynist, anti-feminist movement, that is trying to exert power over women and female-presenting people on the internet. So that's one power dynamic that needs to be taken into consideration. But then there's also the researchers who still hold a little bit of power over even a misogynist, anti-feminist group. And so there are two different [power dynamics] … to be considered. And my research partner and I have taken a lot of inspiration from that because we find ourselves in a similar boat. We are studying a population that is exerting power over other people on the internet, but at the same time, we hold a certain amount of power over them, and we have to weigh both of those as we figure out what we're doing here. (BSR04)

6.8.3. Data curators

Data curators also discussed risk-benefit analysis, especially regarding how published data could potentially harm participants. One curator described internal documentation for assessing harm at the repository where they work, saying, “We have a matrix based on [risk of] harm and [strategies for] deidentification, as to the recommendations we would make to deidentify the data further if it needs it” (DC05). Another curator talked through the tension between informed consent and reproducibility, saying, “We were trying to not only think about consent, but also researchers … [who] wanted to publish the data for reproducibility, for people that are just trying to understand what was going on in research. So we're trying to balance those two things (DC07). Another curator described in detail the various different considerations that come into play when archiving Twitter data:

> I think about what my responsibilities are… to users and to science. I do have a responsibility to Twitter, it just does not trump my other responsibilities. So when I think about what are my responsibilities to the user, I think that when an average user has deleted a tweet that is innocuous and holds little analytic utility, then my obligation is to follow the user's expectation that that tweet will be deleted. But if that would make science harder… So for instance, around the time of the Boston Marathon bombings, Twitter was still quite a popular way for people to respond to crises. Twitter was your real-time social media platform. And [people were] trying to

> identify, where did the bombing occur? Where can people get help? Who are the suspects? Were people searching? Because so many people posted the information that they had at the time, we have an opportunity to study crisis in a way that is not available for other crises that occur. …So [in this case,] our obligation to science and society, I think, outweighs our obligation to any one individual user. (DC08)

One data curator also described talking with colleagues to help them make difficult curation decisions, saying, “Anytime we identify something as a risk, I'll discuss it with my supervisor. And we will develop a plan on how we're going to remediate it” (DC03).

6.9. Data curation issues

The role and process of data curation was a theme that emerged during my deductive coding process. This theme is less concerned with specific data curation strategies, which are included throughout the six key issues above, and instead focuses on identifying the broader benefits, challenges, and concerns relating to data curation. One of the key themes discussed by all three communities of practice—big social researchers, qualitative researchers, and data curators—was the cost and time required to curate data properly (n=10; qr=3, bsr=1, dc=6), and they also talked about their experiences collaborating with curators and repositories to ensure their data were responsibly shared (n=7; qr=4, bsr=1, dc=2). Despite curation-related challenges, participants from all three communities of practice emphasized that the value of big social research and qualitative data sharing made curation efforts worthwhile (n=11; qr=5, bsr=3, dc=3).

Beyond these three areas of overlap, however, the three communities of practice had different ideas and concerns regarding data curation. This may indicate that curation could be an area in which communication between communities of practice could support stronger practices. One subtheme—the concern about the findability of data in official repositories—was mentioned by a qualitative researcher and a big social researcher, but was not mentioned by a data curator (n=2; qr=1, bsr=1, dc=0). However, data curators spoke to every other subtheme. Data curators and qualitative researchers talked about data sharing for the purpose of transparency (n=4, qr=2, bsr=0, dc=2) and suggested that data reuse is

difficult to track, but no big social researchers addressed these ideas. Data curators and big social researchers both talked about data sharing requirements (n=2; qr=0, bsr=1, dc=1) and the technical requirements of big social data and data reuse (n=4; qr=0, bsr=3, dc=1), but no qualitative researchers addressed these ideas. The fact that data curators were able to speak about issues that mattered to big social researchers and issues that mattered to qualitative researchers potentially indicates an ability for data curators to begin to bridge the gap between these two communities of practice.

6.9.1. Qualitative researchers

Several qualitative researchers emphasized the value of qualitative data sharing (n=5). One researcher talked about how data sharing can prevent overburden on participants: “Part of the idea is you're respectful of people's time, don't go ask more people, when you can ask fewer people. Don't ask the same people twice, don't overburden communities” (QR03). Another researcher discussed how data reuse can build on the value of data: “You want people to… use things and adapt [them], you don't just want them to sit on a shelf that nobody ever uses them” (QR06). A third researcher emphasized scientific efficiency, saying, “So many people would not have to [conduct redundant] studies, if we just had the data available (QR08).

Qualitative researchers talked about collaborating with curators and repositories (n=4) in order to support responsible data sharing. One researcher described how a consultation with a data librarian made them feel more comfortable with sharing their qualitative data, saying, “[The data librarian] helped me think of what kind of questions to ask, and so once I felt comfortable with that, with her help I was like, okay” (QR03). Another researcher described using resources from the Qualitative Data Repository (QDR): “We're actually able to refer… students to the QDR’s mechanisms for safely sharing qual[itatitve] data. And that has helped people become compliant with a lot of new NSF mandates. So … QDR has been actually very helpful for that. And has helped I think, in general bolster people wanting to share qualitative data” (QR01).

However, qualitative researchers were also concerned with the cost of data curation—both in terms of money and time (n=3). One researcher who had shared their own qualitative data told me about encouraging their colleagues to do the same, saying, “Nobody does it. They don't take the time. They don't do it. And they're like, "Why should we? what does it give us?" And also people just have demands on their time” (QR03).

Qualitative researchers also noted that qualitative data reuse is rare and hard to track (n=2). One qualitative researcher interviewed other qualitative researchers “about data management, data sharing and data reuse... And what's funny is that none of… the interviewees reused qualitative data” (QR05). But other qualitative researchers talked about sharing data for transparency (n=2), rather than reuse. As one researcher said regarding transparency, “Quantitative computational stuff, it's about try[ing to] get as close as you can to the same results. But for the qualitative stuff, it's more about just making it really transparent. Like, this is what I did. This is why I did it. And this is what I got” (QR07).

6.9.2. Big social researchers

Big social researchers discussed the value of big social research (n=3). One researcher talked about using big social data because of financial constraints: “We need NIH or some [other] type of research grants that many of us in tier two institutions do not have [access to]. In fact the reality is, this is why so many people, including myself, are analyzing social media data in the first place, because we do not have big grants to recruit 1000 people (BSR08). Another researcher talked about the rich and plentiful social interactions that can be pulled from social media, and how those interactions support valuable research outcomes: “[We] use the social media data [to access] this rich, interpersonal textual communication that's happening online, to inform a better understanding of what parts of my community are being stressed are being utilized during some sort of a crisis” (BSR06).

Big social researchers were also concerned with how the technical requirements of big social data can hinder sharing and reuse (n=3). For example, one researcher talked about the difficulty of sharing such a large amount of data, describing a long process of repository selection and negotiation:

> We tried to figure out, where do we put 93 GB [of data]? It… was too big for Zenodo by default, and it was too big for Figshare by default. And so I think we actually contacted Zenodo. And we said, 'Hey, we know that y'all are at CERN and do a bunch of stuff, can we get an exception?' We didn't hear back from them in the time period that we needed. And so we actually went to [our university's institutional repository]. And we even had trouble using [the institutional repository]. So I emailed our data librarian, and our data librarian was like, 'You're not going to be able to upload this to the web interface. But let's work with you.' And everyone was really great in terms of… opening up a back door to upload the 100 GB file. And… I was like, 'Oh, yeah, maybe I actually should've started with y'all at the beginning.' But we've got it there. (BSR06)

Another researcher echoed this sentiment, saying, "GitHub is not good for fairly big datasets, which is what my data… is right now. So I am trying to find a better place to share that dataset. I might just share it on my website as a raw download" (BSR01). This researcher also suggested that they were reluctant to post their data in a data repository, saying, "The issue with uploading stuff on these platforms is they don't show up on search results most of the time. So people won't stumble onto your datasets the way they would on Github. Kaggle is another place where I could upload it" (BSR01). This suggests that datasets in data repositories may be less likely to be found and reused by big social researchers.

6.9.3. Data curators

Data curators were strong believers in the value of data sharing (n=3). One data curator emphasized increased citations as an incentive to publish data: "I try to do a lot of 'Yes, and…' strategies for talking to researchers like, 'Yes, this is great, and…' it will be even more accessible, even more reusable, which means you will get cited more often. I always like to emphasize: if this is reusable, they have to cite you. So if it's more reusable, you'll get more citations" (DC01). Another curator said, "You also shouldn't treat qualitative research as this like, pristine thing that 'Oh, you weren't there. You wouldn't know.' We can still gain value from it (DC04).

But curators also understood the cost and time that is spent preparing data for publication (n=6). One data curator described the time-consuming nature of qualitative data curation: "[I did] my review, and we also have two rounds of quality check on this type of an intensive-level study. So it was roughly 14 weeks of time logged on this study… from assignment [to a curator] to release, which is pretty typical for qualitative [data]" (DC03). Other curators described "the perception [among qualitative researchers] that [data curation] would be time consuming, and [that there wasn't] proper funding for that level of attention" (DC06), and "it's a lot to ask somebody to sit back down and re-transcribe, or even fix automated [transcriptions]. I know how long it's gonna take" (DC09).

To support the value of data sharing, despite curation potentially being time-consuming and costly, data curators talked about how planning for data sharing can make it less of a hurdle (n=4). One data curator said, "My personal interest is in trying to figure out how to get the conversation started with researchers early enough in their research process, so that [data] sharing is not… an afterthought" (DC09). Another data curator shared their strategies for reaching researchers early: "Whenever someone comes to us to ask about, for example, for an NSF project, can you give me a budget, even if they don't ask, we always ask, have you thought about consent for data sharing, because that is a problem. We give workshops. We bring this up all the time, we have templates on our website" (DC02).

In cases where researchers may not have planned ahead for data curation, or when data had other challenges, curators discussed balancing their desire for high-quality data curation with messy reality. Curators told me that "good enough" metadata is sometimes as good as it gets (n=2); for example, a curator said,

> Sometimes it's: gold star, here's a DOI. Like, this is as good as it's going to get. [But] for that we [still] have standards. It's gotta have a good title that's findable [and] someone would be able to recognize the dataset's gist. We need at least a few sentences on what's in the dataset. And ideally, we need a readme, but we're willing to slide on that, [depending on the level of description that is] built into the dataset itself. (DC01)

Another curator said, "[There are] deposit reviews that I've done where PIs [Principal Investigators] have [provided] their coding schemes. It's pretty inconsistent, though, in my

experience with qualitative data, as to when data producers give us that information or not" (DC03).

Curators also thought that sharing some amount of data for transparency purposes was better than sharing nothing (n=2). One curator described a situation in which researchers approached the repository to share their qualitative data, but the consent language the researchers had used with participants didn't allow for sharing: "We're like, you can't just give us the transcripts. It won't fly with your consent [language]. But you could… for all the different codes, themes, notes in your research, [write] a description of your coding strategy, and then [include] one or two extended excerpts [from the interviews]" (DC09).

6.10. Chapter summary

In this chapter, I have provided a detailed view of the results of inductive coding of semi-structured interviews. The results show that qualitative researchers, big social researchers, and data curators all spoke to elements of each of the nine key themes identified in this dissertation: context, data quality & trustworthiness, data comparability, informed consent, privacy & confidentiality, intellectual property, domain differences, strategies for responsible practice, and data curation issues. Participants' views converged or diverged, depending on the themes and subthemes. The interview results are interpreted further in Chapter 7.

Chapter 7. Discussion

This study addressed the following research questions:

RQ1: How is big social data curation similar to and different from qualitative data curation?

> **RQ1a:** How are epistemological, ethical, and legal issues different or similar for qualitative data reuse and big social research?
>
> **RQ1b:** How can data curation practices such as metadata and archiving support and resolve some of these epistemological and ethical issues?

RQ2: What are the implications of these similarities and differences for big social data curation and qualitative data curation, and what can we learn from combining these two conversations?

In this chapter, I discuss how the interviews I conducted with qualitative researchers, big social researchers, and data curators provided insights into these research questions. I begin by outlining the key issues and hypotheses that guided my interviews. The discussion is organized around six key issues—context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. Within the section for each issue, there are subsections that discuss each hypothesis, both suggesting responses to my original hypotheses and discussing new insights. In the final section, I synthesize my research results and discuss the implications for data curation.

7.1. Discussion by hypothesis

When developing the interview guide, I began by outlining hypotheses for each of the six key issues I identified in my literature review. The resulting interview guide included questions that were designed to test each hypothesis. I review each hypothesis below,

including whether the hypothesis was supported, semi-supported, or not supported by the results of my research. An overview of my hypotheses and results is provided in Table 15, on page 163.

7.1.1. Context

In this study, I asked the participants to describe the challenges they encountered relating to preserving, understanding, and communicating the original context in which data were created. Context was one of the most well-thought-out issues among participants. All three communities of practice had considered the question of data context, and had implemented strategies to preserve and communicate context when writing up research and sharing data.

Context hypothesis 1. Qualitative researchers have a stronger concern about context than do big social data researchers

This hypothesis was supported by the results of my research. Each community of practice showed concern about context, but the nature of their concern was distinct for each community of practice, and qualitative researchers' concerns were stronger and more complex than those of big social researchers.

Big social researchers' discussion of context often focused on the more technical aspects of context. These researchers largely discussed the representativeness of social media platforms, the context that could be provided by social media interfaces, and the loss of context that often results from the aggregation of data. However, notably, big social researchers tended to view contextual issues with less concern than qualitative researchers; they acknowledged these issues as a part of the research, but no big social researchers I interviewed thought these issues would compromise their research. Qualitative researchers, on the other hand, were more concerned with how to communicate the deep context inherent in qualitative research—the co-creation of the research, the researcher's background, the community where the study took place, etc. Qualitative researchers were more likely to consider loss of context to be a major obstacle to data sharing. Because qualitative researchers saw the inclusion of contextual information as a vital part of data sharing, they were also concerned about the time required to fully document context.

Qualitative researchers were also concerned that providing details to enhance context could potentially endanger participant privacy and confidentiality.

Context hypothesis 2. Context can be communicated to some extent through embedded or added metadata.

This hypothesis was supported by my research. During my interviews, all three communities of practice discussed metadata and documentation as a potential strategy for preserving context, and data curators were most concerned with how to document context for future use.

> *Context hypothesis 2a. Data curators who specialize in archiving qualitative data can also support metadata that preserves context when archiving big social data.*
> This hypothesis was supported by my research. Data curators discussed the ways in which the inclusion of clear documentation, description, metadata, and related materials alongside shared data could enhance context. Data curators who work with qualitative data described more in-depth review and description processes, and discussed the overlaps between the two types of data. However, data curators were also concerned by the tension between providing enough contextual information for the data to be useful and protecting participant privacy—an issue that applies to both qualitative and big social data.

7.1.2. Data quality and trustworthiness

In this study, I asked the participants to describe challenges they faced relating to data quality and trustworthiness. All three communities of practice discussed documentation, description, and metadata as strategies to support data quality and trustworthiness. All three communities also discussed data completeness as an important element of quality and trustworthiness, especially the need for communicating the level of completeness or missing data. However, each of the three communities offered different ideas about data quality and trustworthiness, and each type of participant emphasized different considerations around data quality and trustworthiness.

Quality hypothesis 1. Big social data is more prone to quality issues than shared/reused qualitative data.

This hypothesis was not supported by the interviews. My research did not show that big social data have more quality issues; rather, these data have different quality issues. The data quality issues discussed in the interviews were wide-ranging and specific to the type of data being analyzed or collected, and each community of practices had unique considerations regarding data quality and trustworthiness. Qualitative researchers were concerned with the human aspects of quality—discussing how data quality was documented in manuscripts, potential researcher bias, trustworthiness in data creators, and the nuances of human communication that are lost when using recordings or transcripts. Big social researchers, on the other hand, tended to focus on technical issues that could affect quality and trustworthiness—spam and bots, programmatic quality issues that arise from computational methods, and including code and related documentation to support quality and trustworthiness.

Quality hypothesis 2. Documentation/metadata can support data quality.

This hypothesis was supported by my interviews with participants. All three communities of practice were concerned with fully describing data quality issues in order to support research integrity and data reuse. All three communities of practice also suggested that when quality issues were well-described in datasets, researchers and curators were more likely to trust that data for reuse.

7.1.3. Data comparability

In this study, I asked the participants to describe challenges relating to comparing and combining different datasets. Participants from all three communities of practice were generally aligned on issues related to data comparability. No qualitative researchers I spoke to had actually compared or combined qualitative datasets, although they agreed on the potential value of this practice for scaling up qualitative results. Some big social researchers had compared and combined datasets, especially to achieve better demographic representation in their research.

Comparability hypothesis 1. Comparing and combining data enables higher quality research (e.g., larger scale, more representative samples, broader conclusions).

This hypothesis was semi-supported by my interviews with participants. All three communities of practice discussed how comparing and combining data can yield stronger research and conclusions. However, only the big social researchers had actually compared or combined datasets. This demonstrates an area in which qualitative researchers' theory is different from their practice. Qualitative researchers could benefit from connecting with big social researchers to support comparability in practice. Comparing and combining datasets is a beneficial strategy to support broader conclusions and more representative samples, and big social researchers' experience with this practice could be applied to support the comparison and combining of qualitative datasets.

Comparability hypothesis 2. Combining datasets is made more difficult for those who reuse qualitative data or use big social data because of challenges relating to missing data, research questions, methods, and metadata interoperability.

This hypothesis was supported by my interviews. While participants understood the theoretical value of comparing and combining datasets, in practice, many were thwarted by challenges that prevent comparability—e.g., data complexity, different data formats, and different metadata formats. Big social researchers were more likely to have successfully combined datasets, especially to support demographic information and more representative study populations. Data curators were especially likely to discuss metadata and format interoperability.

Comparability hypothesis 3. Data comparability issues are similar for qualitative data and big social data.

This hypothesis was supported by my research. All three groups of participants discussed similar issues regarding comparability, regardless of their community of practice.

7.1.4. Informed consent

In this study, I asked the participants to describe challenges relating to informed consent for big social data and archived or reused qualitative data. The issues of informed consent and

privacy overlapped, with many participants discussing deidentification as a strategy for supporting responsible research, even if informed consent was not obtained, or if it is impossible to provide informed consent to unknown future uses.

Consent hypothesis 1. Qualitative and big social researchers have different values and considerations regarding informed consent.

This hypothesis was supported by the interviews. The issue of informed consent produced the widest range of responses among the participants. All three communities of practice touched on the role of the Institutional Review Board (IRB), but most emphasized that the IRB was not a helpful resource for issues of data sharing and reuse. Participants described how IRB protocols are not designed to regulate data reuse or big social data, and they noted that the heterogeneity of IRBs at different institutions resulted in researchers receiving different or inconsistent guidance from different IRBs. Other than topics relating to IRBs, the concerns of qualitative researchers and big social researchers regarding informed consent did not overlap. (See *Consent Hypotheses 1a and 1b* for further discussion of this difference.) Data curators weighed in on most themes, supporting the idea of data curators as well-positioned to build connections between communities of practice.

> *Consent hypothesis 1a. Qualitative researchers are more strict about issues related to consent and reuse, even for archived data/data reuse.*
>
> This hypothesis was supported by the interviews. Qualitative researchers were generally uncomfortable with the idea of research participants consenting to future use of data. Many qualitative researchers whom I spoke to had included data sharing in their consent forms, including using tiered consent models. (See Chapter 2, page 30 for a discussion of tiered consent.) However, qualitative researchers still had concerns about whether research participants fully understood the potential future uses of the data and the potential risks of that reuse. Some qualitative researchers used restricted access maintaining participant privacy. Big social researchers were generally not concerned about obtaining the consent of the people whose data was collected through their research, and they did not consider this necessary. See below, *Consent hypothesis 1b,* for more detail.

Consent hypothesis 1b. Big social researchers are more open to using creative strategies to address consent (e.g., focus groups, community advisory groups).
This hypothesis was not supported by the interviews. Big social researchers did not report using consent strategies such as focus groups or community advisory groups. Some participants had considered the problematic nature of consent for big social data, and some of them described designing their research to reduce the potential harm for participants who had not explicitly consented to the research. However, some big social researchers told me that they did not consider their research to be human subjects research at all, and that informed consent was therefore unnecessary. None of the big social researchers I interviewed had taken steps to obtain participant consent beyond the blanket user agreement in social media platform terms of service.

Consent hypothesis 2. Qualitative data curation approaches for consent could be adapted to fit big data researchers.
This hypothesis was semi-supported by the interviews. My research shows that most big social researchers do not obtain consent from participants. Big social researchers generally consider the consent that users provide when agreeing to social media terms of service to be sufficient, and the norms and values of the big social research community do not require going further to obtain additional consent. Unless community norms change and big social researchers conclude that informed consent is ethically necessary for their research, it is unlikely that big social researchers will adopt data curation strategies that support informed consent.

My research suggests that community norms and ethical standards differ significantly between the qualitative research community and the big social research community. In qualitative research, those norms and standards require that participants specifically consent to data sharing and data reuse, whereas community norms and standards in the big social research community do not require participants' consent. My research suggests that curators can use a few strategies to protect participants even if informed consent was not obtained: ensure deidentification of data, provide restricted access, or provide data enclaves where reusers can analyze big social data without downloading it. Data curators have the

expertise and perspective to help researchers responsibly use and reuse data—for example, by considering the sensitivity of data and engaging with specific issues on a case-by-case basis. While curators generally deferred to researchers as the experts in their own domains, curators did have a strong sense of ethical responsibility toward social media users and qualitative research participants, and they had concerns about the lack of informed consent in big social research. Curators discussed the importance of connecting with researchers early in the research process as the key strategy for supporting consent. At this early stage, curators could encourage creative consent practices such as a participant opt-in for big social research studies, or the use of community focus groups or community advisory groups, if applicable.

7.1.5. Privacy and confidentiality

In this study, I asked the participants to describe challenges relating to privacy and confidentiality of research participants, including the people represented in big social data. Among the three communities of practice, there was consistency in how participants understood and addressed the issue of privacy and confidentiality. Qualitative researchers, big social researchers, and data curators all had a similar level of concern about privacy and confidentiality. All three communities of practice had considered research practices relating to data deidentification, data sensitivity, restricted access, participant/user expectations of privacy, potential harms to participants, research design for privacy, and data security.

Privacy hypothesis 1. Big social data researchers are less concerned about privacy than qualitative researchers.

This hypothesis was not supported by the interviews. Even though big social researchers did not consider it necessary to obtain informed consent, those I interviewed were highly concerned about participant privacy. Big social researchers told me about a variety of strategies for protecting privacy and confidentiality, and one participant described their internal struggle about their research being part of a broader system of online surveillance. Also, most big social researchers I spoke to did not share their big social datasets with others, due to concerns about participant privacy and social media companies' intellectual property rights. (See also section 7.1.6. Intellectual property and data ownership.) Instead,

when sharing research materials, big social researchers were more likely to provide the code that would allow potential future researchers to reproduce the original research by collecting the data themselves.

Privacy hypothesis 2. Data curation practices for supporting privacy with qualitative data can inform big social data.

This hypothesis was supported by the research. Because of the similarity of privacy and confidentiality related issues among the three communities of practice, curation strategies for protecting privacy were also applicable to both qualitative data and big social data. Curation practices that were highlighted in my interviews included deidentification, restricted access, and applying different levels of curation depending on the sensitivity of the data.

7.1.6. Intellectual property and data ownership

In this study, I asked the participants to describe challenges relating to intellectual property and data ownership. Participants generally had limited understandings of intellectual property (IP) and data ownership, and few had considered these issues in detail. To address issues of intellectual property and data ownership, participants discussed purchasing or using commercially-available data, data licensing, data citation, and directly contacting participants or organizations involved in the original research.

IP hypothesis 1. IP is a more important issue for big social researchers than researchers who reuse qualitative data.

This hypothesis was supported by my research. Most qualitative researchers had not considered the intellectual property rights or data ownership of research participants. One qualitative researcher who had reused decades-old ethnographic data discussed concerns about data ownership, especially for indigenous research participants. Some qualitative researchers considered the ideas of data citation and data licensing for their own data. However, for most qualitative researchers, IP concerns did not greatly affect their practices of data sharing and reuse.

IP hypothesis 2. IP concerns may prevent big social data researchers from archiving data.
This hypothesis was supported by my research. While a few big social researchers whom I spoke with described breaking social media terms of service, most big social researchers felt obligated to adhere to the terms of service regarding big social data use. The majority of big social researchers I spoke to had not made their research data publicly available, opting instead to describe their methods for collecting the data so that future researchers could replicate the data collection process for themselves.

IP hypothesis 3. Data curation practices that address IP issues for qualitative data can inform big social data and vice versa.
This hypothesis was supported by the research. Data curation practices such as data licensing and encouraging data citation can support intellectual property rights and data ownership concerns for both qualitative and big social data. Data repositories and data curators can also review data to ensure that IP rights and data ownership concerns are addressed. For big social data, repositories can provide tools such as hydrators for Tweet IDs or instructions and code for how big social data can be re-collected; these tools allow researchers to share some elements of their data while complying with the applicable terms of service. Data repositories can also address intellectual property concerns by restricting use of the data to those who meet certain conditions, or by providing analytical outputs rather than sharing a full dataset.

Table 15. Overview of hypotheses and results by issue

Issue	Hypothesis	Result
Context	1. Qualitative researchers have a stronger concern about context than do big social data researchers	Supported. While both communities were concerned with context, qualitative researchers saw context as a much more complex issue than did big social researchers.
	2. Context can be communicated to some extent through embedded or added metadata.	Supported. All three communities of practice discussed this idea.

	a. Data curators who specialize in archiving qualitative data can also support metadata that preserves context for big social research and big social data.	Supported. Strategies for communicating context are similar.
Data quality and trust-worthiness	1. Big social data is more prone to quality issues than shared/reused qualitative data.	Not supported. Both big social data and qualitative data may have quality issues. However, the nature of these issues are different for each type of data.
	2. Documentation/metadata can support data quality.	Supported. All three communities of practice discussed this idea.
Data comparability	1. Comparing and combining data enables higher quality research (e.g., larger scale, more representative samples, broader conclusions).	Supported. Combining datasets was rare in practice, but more common in big social research.
	2. Combining datasets is made more difficult for those who reuse qualitative data or use big social data because of challenges relating to missing data, research questions, methods, and metadata interoperability.	Supported. These data comparability challenges are common across communities of practice.
	3. Data comparability issues are similar for qualitative data and big social data.	Supported, although few qualitative researchers had compared or combined data.
Informed consent	1. Qualitative and big social researchers have different values and considerations regarding informed consent.	Supported. These two communities of practice had different viewpoints on consent.
	a. Qualitative researchers are more strict about issues related to consent and reuse, even for archived data/data reuse.	Supported. Qualitative researchers were cautious about issues of consent.
	b. Big social researchers are more open to using creative strategies to address consent (e.g., focus groups, community advisory groups).	Not supported. Big social researchers did not report using consent strategies beyond terms of service.

	2. Qualitative data curation approaches for consent could be adapted to fit big data researchers.	Semi-supported. Community norms regarding informed consent are starkly different for qualitative researchers and big social researchers. However, curation can protect participants from harm, even if informed consent was not obtained.
Privacy and confidentiality	1. Big social data researchers are less concerned about privacy than qualitative researchers.	Not supported. Even when big social researchers were not concerned with consent, they were concerned with privacy.
	2. Data curation practices for supporting privacy with qualitative data can inform big social data.	Supported. Big social researchers look to data curators and existing strategies such as deidentification and access control to protect the privacy of research subjects.
Intellectual property and data ownership	1. IP is a more important issue for big social researchers than researchers who reuse qualitative data.	Supported. More IP issues are present with privately- controlled big social data.
	2. IP concerns may prevent big social data researchers from archiving data.	Supported. Few big social researchers had shared their data, and IP concerns were a major influencing factor.
	3. Data curation practices that address IP issues for qualitative data can inform big social data and vice versa.	Supported. Similar strategies can be used for both data types.

7.2. Synthesis

Three analytically powerful themes emerged through deductive coding: domain differences, strategies for ethical, legal, and epistemologically-sound research (referred to in shorthand as “responsible research”), and data curation issues. I also identified two overarching ideas that were present across all codes: human subjects vs. content and different focuses and approaches for each issue. These themes respond directly to my research questions by highlighting the similarities and differences between the communities of practice, and by outlining how data curation strategies can support connecting communities and building

standardized practices that support responsible practice for both qualitative data reuse and big social research.

7.2.1. Domain differences

The participant group from each community of practice included a variety of disciplines (see Chapter 5, Table 11 and Table 12). I found that disciplinary norms and values are key factors that affect how each community of practice addresses issues. Each community of practice's approach, and their prioritization of key issues, varied according to the specific discipline within that community of practice. Despite disciplinary differences, researchers were unified within their community of practice due to the data types they worked with and their research methodologies.

Different communities of practice approached risk-benefit analyses differently; they came from different backgrounds and had different values and priorities, so their analyses had different outcomes. For example, qualitative researchers focused more on the human participants underlying the qualitative data, and they therefore valued participant consent and privacy above other potential benefits. Big social researchers tended to abstract their data from the human participants who created them, focusing instead on technical considerations.

Different communities of practice also tended to have different skills, training, and background. One big social researcher described how rare it is to find researchers who have both the technical skills for computational data collection and analysis, and training in social science ideas and methodologies. As this researcher said, "the Venn diagram of the people who can do [both] … is vanishingly small" (BSR06). With this in mind, both qualitative researchers and big social researchers talked about the idea of looking to other disciplines for inspiration and collaborating with other disciplines to support scaled-up, responsible research. However, few participants reported specific instances of connecting with researchers from other disciplines or communities of practice—and for those who did, the researchers from other communities of practice tended not to be full collaborators, but instead served in a consultant role.

7.2.2. Strategies for responsible practice

The participants I interviewed often drew on many sources to cobble together strategies for responsible practice—that is, epistemologically-sound, ethical, and legal practice. Researchers described a process of continuous re-examination of epistemological, ethical, and legal issues—making decisions on the fly about how to act responsibly. Researchers used several strategies for decision-making and problem-solving to support responsible practice: informal risk-benefit analyses, thinking through challenges on their own, talking to colleagues and collaborators, reading the literature, and implementing strategies they had learned in graduate school.

Participants usually talked with an IRB or obtained exempt status from an IRB, but as a general rule, IRBs do not review research that uses existing data, whether that be data reuse or big social research. It was rare for participants to discuss ethical guidelines or standard community best practices. Only two researchers referred to community standards, and only one referred to the Association of Internet Researchers (AoIR) Ethical Guidelines. This may be related to disciplinary silos. Social scientists reusing qualitative data would likely not consider looking to the AoIR for guidance on data reuse—and, in fact, the AoIR ethical guidelines are designed for big social researchers, not qualitative data sharers or reusers. The big social researchers I talked to had a variety of disciplinary backgrounds (civil engineering, communication, computer science, information science, journalism, public health), but no participants reported that their academic training included responsible big social research practices. It is possible that the researchers misreported their level of training—that they simply failed to retain the information they were taught in graduate school on this subject. Alternatively, if the researchers were accurately reporting a lack of instruction on this subject, academic training may begin to address these issues in more detail as big social research grows more common in these disciplines.

A key takeaway from this research is that all three communities valued responsible research practices, but most did not have clear training on these practices or resources to turn to. Because IRBs do not review research that uses existing data, researchers who use such

data—including big social researchers and those who reuse qualitative data—cannot rely on IRBs to provide ethical guidance, and they are left to fend for themselves. Researchers and curators from all three communities would benefit from concrete guidelines, ethical codes, and tools or workflows that support risk-benefit analysis and harm reduction.

7.2.3. Data curation issues

During their interviews, participants often discussed the broad benefits and significant challenges of data curation. While a high number of participants talked about the value of data sharing, many also pointed to the time-consuming nature of data curation. And data curation becomes all the more time-consuming and complex if curators and researchers aim to fully address issues of context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. Still, many participants discussed their successful experiences collaborating with data curators to support data sharing and reuse.

Qualitative researchers and big social researchers generally had different ideas and concerns regarding data curation. Qualitative researchers were concerned with transparency rather than reproducibility or reuse, pointing out that qualitative data reuse is rare. Big social researchers were concerned about how data curation could support technical considerations such as compliance with data providers' terms of service, computational methods, and software dependencies. Knowing that these two groups of researchers focus on different considerations can help data curators better serve these communities of practice. Data curators were concerned with how to encourage researchers to share data, despite the time and effort required. Here, too, understanding researchers' different needs and priorities can enable better advocacy by data curators and can also support tailored data curation resources that respond directly to researchers' different needs.

Data curators as a community of practice discussed almost all of the data curation-related subthemes.[4] This indicates both that data curation is an area in which communication

[4] Some researchers pointed out that data in repositories are often not findable with a Google search, and one suggested that they were more likely to post data on a personal website for that reason. No data curators discussed this topic, although it is an important insight for data repositories looking to promote discovery and reuse of archived data.

between communities of practice could support stronger practices, and that data curators could act as a bridge between qualitative researchers and big social researchers.

7.2.4. Human subjects versus content

Qualitative researchers and big social researchers demonstrated a striking difference in approach regarding what constitutes "human subjects" data. Qualitative researchers were deeply considerate of human subjects, focusing on the participants as co-creators who were giving the gift of their experience to the research process. Big social researchers, on the other hand, focused on technical considerations, and were more likely to think of big social data as unembodied "content," rather than as an extension of the human participants who created the content. This foundational philosophical mismatch between qualitative researchers and big social researchers provides insight into key differences between the two communities' approaches to the issues of context and consent, although it did not affect participants' approaches to privacy.

When discussing context, big social researchers had no expectation that a complex view of context would be available to them. They were concerned with technical considerations such as how the data were presented when collected through an API versus an online interface, or embedded metadata indicating time, location, user bio, hashtags, or conversation threads. Qualitative researchers viewed context differently. They discussed the difficulty of understanding the entirety of a participant's personal background; they considered nuances such as tone of voice, gestures, and body language; and they highlighted the idea of researcher as a contextual factor, including how the relationship between researcher and participant affected results. Some big social researchers did combine multiple datasets to enhance context by incorporating demographic information or providing additional viewpoints. Big social researchers were more likely to compare and combine datasets to enhance context and create larger, more complex datasets. Data curators could encourage qualitative researchers to use this strategy as well. Comparing and combining datasets can support scaling up qualitative research, and could potentially be one strategy to enhance the context and meaning of reused data.

The philosophical divide as to whether reused data should be viewed as human subjects data was also an important factor when discussing informed consent. As noted above in section 7.1.4., informed consent for data reuse was a major issue for qualitative researchers, and qualitative researchers were concerned about participant consent for research with archived or reused data; qualitative researchers considered archived data to still be human subjects data. On the other hand, the big social research community has adopted the view that collecting content from online sources is not human subjects research and can therefore be done freely, without user consent. If there was concern voiced by big social researchers, it was generally regarding the terms of service or the intellectual property rights of the private companies that make big social data available.

However, regarding privacy, both big social researchers and qualitative researchers were aligned, and all three communities of practice used similar data curation strategies to ensure participant privacy. These strategies included deidentification, restricted access or data enclaves, and research design to support privacy. Big social researchers' recognition of the value of privacy may be an opportunity for data curators to engage with big social researchers. And valuing user privacy is a first step toward suggesting to big social researchers that their data should be viewed as human subjects data. Data curators can provide strategies that support privacy, while also providing guidance relating to context and consent.

7.2.5. Different focuses and approaches for each issue

While the interviews showed that six key issues (context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership) were applicable to all three communities of practice, each community of practice had different focuses and approaches for each issue. Due to their different backgrounds, training, data types, and values, qualitative researchers, big social researchers, and data curators saw each issue through a slightly different lens. As noted above, qualitative researchers were trained to focus on, and analyze, how the complexities of participants' (and researchers') life experience and perspectives can affect the data and the data analysis. On the other hand, big social researchers were accustomed to the idea that

social media posts and other big social data lack the full contextual details of a person's life; instead, big social researchers focused on understanding social media platforms, code, technologies, and demographics. Data curators brought a third approach to the issue of context, based upon their foundation of training in metadata, documentation, and preservation; data curators were most focused on how to communicate context to future users, and how to provide access to data in its original context whenever possible. This variety of different focuses and approaches was apparent in the participants' discussion of each of the six key issues, and it is also reflected in the four synthesis sections above—domain differences, strategies for responsible practice, data curation issues, and human subjects versus content.

These different focuses and approaches demonstrate the value of connecting the three communities of practice. Each community discussed different aspects of each issue, but all of these aspects can be applied across each data type. As qualitative data sharing and reuse grows, qualitative researchers will benefit from considering the focuses and approaches that were discussed by big social researchers. Similarly, as big social researchers increasingly consider the epistemological, ethical, and legal complexity of big social research and big social data sharing, they will benefit from considering the focuses and approaches that were discussed by qualitative researchers. Data curators, for their part, must not only be able to understand epistemological, ethical, and legal issues from the perspective of post-research documentation and preservation, but also be aware of the complexities that arise during the research process, prior to the data sharing stage. In the interviews, data curators were aware of the benefit of discussing data curation with researchers early in the research process; this is discussed further below. I provide an overview of the similarities and differences among the three communities of practice in Table 16 and Table 17. Table 16 lists the focuses and approaches that were discussed and assigned similar importance by all three communities of practice. Table 17 lists the focuses and approaches that were distinct between each of the three communities of practice.

Table 16. Similar focuses and approaches between the three communities of practice, according to issue

Issue	Similar focuses and approaches between the three communities of practice
Context	● Documentation, description, and metadata ○ Fully describing population, research conditions, researcher and participant details ○ Providing related materials ● Time and resources necessary to create thorough documentation, description, and metadata ● Tension between providing contextual details and retaining participant privacy
Data quality and trust-worthiness	● Documentation, description, and metadata ○ Fully describing data quality issues supports data reuse ○ Datasets with a clear explanation of quality issues were seen as more trustworthy ● Data completeness as an important element of quality and trustworthiness—identifying what data was used in the analysis, what data is archived, what data may be missing.
Data compara-bility	● Challenges that hinder comparing and combining datasets: data formats, metadata standards, language, encoding language ● Benefits of comparing and combining data—more data could lead to stronger research conclusions ● Documentation and metadata can support comparability and interoperability
Informed consent	● Considered IRBs' role in informed consent procedures, but had different ideas about the role of the IRB.
Privacy and confid-entiality	● The issue of privacy and confidentiality had the most consistency between the three communities of practice. Privacy and confidentiality considerations: ○ Data deidentification ○ Taking more care with vulnerable or sensitive populations ○ Access controls to support privacy ○ Participant/user expectations of privacy, depending on data source and other contextual factors ○ Consideration of potential harms ○ How research design can support privacy ○ Data security concerns
Intellectual property	● Lack of clarity about IP rights and data ownership, including some hesitancy among participants to speak about IP, citing lack of expertise. Other considerations: ○ Purchasing/using commercially-available data to clarify IP and data ownership concerns ○ Data licensing ○ Data citation ○ A rare, but notable, practice was to contact participants and organizations involved in the original research to discuss data reuse.

Table 17. Focuses and approaches that were different between the three communities of practice

Issue	Different focuses and approaches addressed by each community of practice		
	Qualitative researchers	**Big social researchers**	**Data curators**
Context	Focused on communicating the complex context of qualitative research: ● Co-creation of research ● Background of researcher and participants	Focused on technical aspects of context: ● Representativeness of social media platforms ● Challenge of context and aggregated data ● Context provided by online interfaces	Focused on documentation: ● Metadata ● Readmes ● Links to related materials
Data quality and trust-worthiness	More likely to include discussion of data quality and trustworthiness in the manuscript, not with the dataset. Other unique focuses: ● Quality of transcripts, videos, and recordings ● Researcher bias ● Outdated research practices could lead to lower quality or less trustworthy data	More likely to include documentation of data quality with the dataset, especially related code. Other unique focuses: ● Representativeness of dataset affects quality ● Issues with large-scale data collection and automation—spam, bots, programmatic issues ● Combining datasets to support quality	Generally considered data quality to be outside of their purview, instead focusing on the quality of metadata and documentation, as well as repository trustworthiness. Other unique focuses: ● Ensuring that data can be readable, code is executable ● Curator review supports quality and trust
Data compar-ability	Had not considered data comparability as thoroughly as other communities of practice. Unique focuses: ● Designing studies for combined use ● Complexity and flexibility of qualitative methods make comparing and combining qualitative data difficult	Had combined datasets more often, to support broader conclusions and broader study populations. Unique focuses: ● Challenges of matching different datasets ● Interoperability challenges—language, metadata, data formats	Were concerned with metadata and format interoperability. Unique focuses: ● Documentation and training ● Standardized metadata schemas (e.g., DDI, DataCite) ● Nonproprietary file types ● Interoperability between qualitative data analysis systems (e.g., NVivo, Atlas)

Informed consent	Deeply concerned with how to responsibly maintain informed consent for shared data. Unique focuses: ● Skeptical of IRBs' ability to support ethical data sharing and reuse ● Consent language and procedures ● Allowing participants to review and redact transcripts ● Concern about unknowns: future use of data, participants' understanding of consent procedures	Generally not concerned with informed consent. Unique focuses: ● IRB involved only if research affects human behavior (e.g., altering social media timelines or adjusting reputation score) ● Data considered public content, not human subjects data ● Efforts to design research so consent was less important ● Deidentification as a strategy to protect users, even consent is unclear	Generally concerned with responsibility of data repository, while providing access to shared data whenever as possible. Unique focuses: ● Collaborating with IRBs to support consent procedures ● Ensuring consent was given for data sharing ● Facilitating as much sharing as possible if no consent was given. ● Deidentification as a strategy to protect participants if consent is unclear
Privacy and confiden-tiality	Well-established concerns and strategies. Unique focuses: ● Deidentification of full transcripts ● Weighing potential harms to participants	Concerns and strategies were less established, more ad hoc. Unique focuses: ● Deidentification of quotes in published articles ● Research design to support privacy	Concerns related to repository and curator support for privacy. Unique focuses: ● Curator support for deidentification ● Levels of restricted access
Intellectual property	More concerned with people than institutions. Unique focuses: ● Data sovereignty for participant communities ● Data citation to support IP rights and data ownership	More concerned with companies and institutions. Unique focus: ● Complexities of social media terms of service	Concerned with IP as it relates to data repositories. Unique focuses: ● Repository terms of use ● How data sovereignty ownership, and governance affect shareability

7.3. Implications for data curation practice

My research questions asked: How is big social data curation similar to and different from qualitative data curation? What are the implications of these similarities and differences for big social data curation and qualitative data curation, and what can we learn from combining these two conversations? The results of my research suggest that data curation strategies

can support and enhance responsible practice in some cases, and that data curators can act as facilitators and intermediaries between communities of practice.

Data curators were able to speak fluently about a variety of issues—both those that concerned big social researchers and those that concerned qualitative researchers. This indicates that data curators have the ability to begin to bridge the gap between these two other communities of practice, and to mediate and translate the different requirements and perspectives of each community of practice. Especially when they were able to consult with researchers throughout the research lifecycle, data curators were able to observe a broad range of the issues confronting both qualitative researchers and big social researchers, and to evaluate the communities' focuses and approaches for those issues.

Participants also suggested specific strategies for data curation relating to the six key issues. As an example, intellectual property was confusing to everyone. Participants were relatively unsure about what IP law meant and how it impacted their research, but they were aware of how data curation could support IP rights, especially data curation-related strategies such as data citation, data licensing, and restricted access.

Other data curation strategies included help with deidentification and help with metadata and description, including standardized metadata and file formats to support interoperability. Curators can review consent forms prior to research, ensuring that consent to data sharing is clear. Curators can also request and review materials such as interview guides, software, and code; these related materials may be included as part of a data deposit to mitigate epistemological issues. Table 18 provides an overview of each key issue, the aspects of that issue addressed by data curators in their interviews, and the applicable data curation strategies that curators use to address each issue.

Table 18. Aspects of issues addressed by data curators and coinciding data curation strategies

Issue	Data curator focuses	Data curation strategies
Context	Focused on documentation: ● Metadata ● Readmes ● Links to related materials	● Work with researchers to include in-depth documentation, metadata, and linked materials alongside datasets in repositories
Data quality and trust-worthiness	Felt that data quality was outside of their scope, instead focusing on the quality of metadata and documentation, as well as repository trustworthiness. Other focuses: ● Ensuring that data can be readable, code is executable ● Curator review supports quality and trust	● Work with researchers to support thorough, high-quality metadata and documentation ● Pursue CoreTrustSeal certifications for repositories and/or align with TRUST Principles. ● Check data and code to ensure it is readable and executable.
Data compar-ability	Were concerned with metadata and format interoperability. Focuses: ● Documentation and training ● Standardized metadata schemas (e.g., DDI, QuDEx, DataCite) ● Non-proprietary file types ● Interoperability between qualitative data analysis systems (e.g., NVivo, Atlas)	● Provide documentation and training for researchers to support comparing and combining data ● Use standardized metadata whenever possible ● Provide training and guidance on metadata standards, non-proprietary file types, and open source software ● Continued advocacy for interoperability between qualitative data analysis systems
Informed consent	Generally concerned with responsibility of data repository, while providing access to shared data whenever as possible. Focuses: ● Collaborating with IRBs to support consent procedures ● Ensuring consent was given for data sharing ● Facilitating as much sharing as possible if no consent was given. ● Deidentification as a strategy to protect participants if consent is unclear	● Collaborate with IRBs, research offices, etc. to support consent procedures early in the research process ● Point researchers to appropriate resources such as domain-specific codes of ethics ● Curatorial review of data for sharing, to ensure consent was appropriate ● Support and training for deidentification ● Facilitating partial sharing for transparency if consent procedures do not allow full data sharing ● Restricted/controlled access for shared data

Privacy and confiden-tiality	Concerns related to repository and curator support for privacy. Focuses: ● Curator support for deidentification ● Levels of restricted access	● Support and training for deidentification ● Restricted/controlled access for shared data ● Point researchers to appropriate resources such as domain-specific codes of ethics
Intellectual property	Concerned with IP as it relates to data repositories. Unique focuses: ● Repository terms of use ● How data sovereignty ownership, and governance affect shareability ● Data citation ● Data licensing	● Training for researchers on IP concepts ● Repository terms of use ● Data citation ● Data licensing ● Rights clearance and management for reused datasets

Data curators emphasized the importance of planning ahead for responsible big social research, data reuse, and data sharing, but they also told me that it is difficult to reach researchers early in the research process. Researchers often approached curators only after their research was complete, rather than at the outset of a project. My research suggests two potential solutions to this challenge. First, data curators can use collaborations to support early contact with researchers. IRBs, research support offices at universities, and big data providers could all be potential partners for data curators, helping to bring in data curators earlier in the research lifecycle. Second, by documenting the concerns and issues of big social researchers and qualitative researchers, my research identifies areas of concern that can function as entry points for data curators to connect to researchers. Data curators can offer services specifically tailored to the issues and concerns identified by this research, such as review of consent procedures to support data reuse, review of social media terms of service, or review of big social research design, with an eye toward epistemologically-sound, ethical, and legal practice.

As data curators build relationships with qualitative and big social researchers, they can also act as translators and knowledge brokers to support interconnection among the communities of practice. The issues identified in my dissertation are continually being examined, and codes of ethics and other guidelines for responsible practice are still being developed. Because data curators' knowledge of data curation spans different domains and

disciplines, data curators are well-situated to be advocates for responsible practices relating to data use, sharing, and reuse.

7.4. Chapter summary

This research shows that qualitative researchers and big social researchers, as distinct communities of practice, are indeed under-connected. While participants indicated that they did look to other disciplines and domains for inspiration or guidance, it was rare for colleagues from other domains to be included as full collaborators in a research team. My research also suggests that data curators can support connections between the communities of practice. Data curators had extensive experience with and a ready understanding of a variety of issues due to their working relationships with both big social researchers and qualitative researchers, as well as their experience curating both big social data and qualitative data. This indicates that there is an opportunity for data curators to build connections between these two other communities of practice.

Qualitative researchers and big social researchers both viewed data curation as time-consuming but potentially helpful. However, researchers were not aware of all of the ways in which data curators and repositories are available to support responsible research practices. Even though my research demonstrates that data curators are aware of many of the issues confronting researchers and can suggest strategies for responsible practice, it was rare for researchers to contact data curators before their research was complete. Researchers usually viewed data sharing as a final step in the research process, and they did not interact with data curators until they began the data sharing process in a data repository. Instead of relying on the experience and advice of data curators, qualitative researchers and big social researchers relied on informal strategies to support responsible practice, including informal, iterative risk-benefit analyses, conversations with colleagues, reading the relevant literature, ethics training, and peer review feedback on publications.

This research suggests that data curators have a broader view of all disciplines, domains and methodologies, and are therefore well-positioned to help build bridges between the communities of practice and support responsible practice in big social research and

qualitative data reuse, using the strategies outlined in section 7.3. However, data curators as a community of practice are also under-connected with qualitative researchers and big social researchers. Encouraging connection between all three of these communities of practice will support more responsible research, as well as enhanced and increased data sharing, thus leading to additional discoveries and insights in behavioral and social science.

Chapter 8. Conclusion

This dissertation has examined the similarities and differences between qualitative data reuse and big social research, and how data curation practices could address epistemological, ethical, and legal issues presented by these two types of research. In Chapters 2, 3, and 4, I identified key issues common to qualitative data reuse and big social research, and then I outlined data curation strategies that could alleviate some aspects of these issues. In Chapter 5, I described the theoretical framework for my research, including how the theories of community of practice and epistemic cultures informed my research. I then explained how I identified the six key issues through a review of the literature, and I detailed my process of conducting semi-structured interviews using critical incident technique to learn how qualitative researchers, big social researchers, and data curators address each of these issues in practice. In Chapter 6, I detailed the results of a qualitative content analysis of the interviews, using grounded theory's constant comparison method. In Chapter 7, I discussed the implications of these results for data curation practice.

8.1. Contributions

This dissertation shows that the three communities of practice I investigate—qualitative researchers, big social researchers, and data curators—are under-connected. My research shows that all three communities of practice are affected by the same six key issues when conducting big social research, qualitative data sharing and reuse, and data curation, and that the three communities of practice often emphasize different aspects of those issues.

The fact that the members of each community of practice often spoke to different aspects of each issue is precisely why it would be beneficial for these three communities to come together. The different aspects of these issues are key to connecting research communities and data curators for their mutual benefit: the focuses and approaches that are emphasized by each individual community could potentially benefit the other communities. Each community can learn from the other, especially to identify and consider aspects of these issues that might not naturally occur to them.

The issue of informed consent is an illustrative example. Big social researchers' community norms dictate that the use of big social data does not require informed consent for each specific research project. Most big social researchers consider the broad consent that users provide when signing up for online services to be a sufficient level of consent for all aspects of big social research. Because big social researchers often work without specifically informed consent, they have developed other strategies to reduce potential harms to participants—for example, they consider what research questions will provide important insights without posing undue risk to participants, they are careful to deidentify direct quotes, and they use strategic data-sharing strategies such as restricted access or sharing TweetIDs that must be rehydrated by future users. However, qualitative researchers are accustomed to one-on-one interactions with participants, including obtaining careful informed consent for each research study. This causes some cognitive dissonance when considering alternative strategies for consent that facilitate data sharing and reuse. As qualitative data sharing becomes more common, qualitative researchers may benefit from adapting some of the strategies that big social researchers use to protect participants, even if truly informed consent may be impossible. These strategies can help qualitative researchers realize the benefits of qualitative data reuse to scale up qualitative research and build longitudinal studies that enhance discoveries in social and behavioral science.

On the other hand, qualitative researchers' consideration of the human element of archived and big social data could be a beneficial lens through which big social researchers could view their research, encouraging big social researchers to take even more care when considering ethical issues, and providing a more nuanced perspective of epistemological issues. Again using the example of informed consent, qualitative researchers could help balance big social researchers' ideas about consent, encouraging big social researchers to consider strategies for automatically obtaining consent from social media users and primary research subjects, or alternative strategies for consent such as talking with community focus groups about the research. These additional considerations relating to consent could potentially expand big social researchers' ability to responsibly study vulnerable populations and sensitive topics.

This research also shows that data curators as a community of practice lack sufficient connection to qualitative researchers and big social researchers. Many qualitative and big

social researchers whom I interviewed were unaware of the extent to which data curators could collaborate with and assist them to support responsible data practices. Data curators' services and skills are therefore under-used.

The data curators interviewed in my study had thought deeply about data reuse and big social research, and they were therefore familiar with a variety of issues affecting these two types of research. Also, despite the different aspects of each issue that were discussed by qualitative researchers and big social researchers, the data curation strategies for these types of research were often similar. Metadata, description, nonproprietary file formats, open source software, permanent identifiers, access controls, and links between related materials are all data curation strategies that support the six key issues identified in this research—context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. Data curators are well-positioned not only to act as curation experts and repository managers, but also as community connectors and translators, facilitating connection between qualitative researchers and big social researchers through their broad knowledge of data curation for both communities.

The qualitative researchers and big social researchers who were interviewed for my dissertation research rarely contacted data curators before their research was complete and they were actively considering sharing their data. This meant that the researchers were not able to benefit from data curators' broad knowledge during the research process; instead, they cobbled together informal strategies to support responsible practice. This dissertation suggests that data curators should focus on connecting with researchers early in the research process—through partnerships with IRBs, university research support offices, and big data providers. By describing issues of particular concern to big social researchers and qualitative researchers, this dissertation also highlights areas in which data curators can offer specific services—for example, data curators can provide reviews for consent procedures that support data reuse, social media terms of service, or big social research design.

8.2. Limitations

This study had a few limitations. First, the study was relatively small, and the sample of participants was influenced by volunteer bias. Information scientists were more likely to respond to my request for participation due to their affinity to the research topic. Others who volunteered may have done so because they had worked with data curators and librarians in the past, and therefore felt that they wanted to contribute to the field.

Second, I usually asked the questions in the same order for each interview. Purposefully randomizing the order of the issues addressed could potentially have strengthened the study, because participants may have higher energy levels at the beginning or middle of an interview. By switching the order of issues so that they were raised randomly throughout the interview, I might have helped control for the varying levels of engagement over the length of each participant's interview.

Third, my skill as a researcher improved throughout the interview and analysis process. The more interviews I conducted, the more my ability improved to ask follow-up questions, prompt participants for additional information, and generally guide participants through the interview process. Therefore, it is possible that the interviews I conducted later in the process elicited more in-depth information from those participants, and that my earlier interviews were not as successful in eliciting all the information that the participants might have provided. During a few of my first interviews, there were moments where participants strayed from explaining or analyzing their critical incident and began speaking in generalities or getting off-topic. While all 30 interviews ultimately produced useful data, my ability to guide interviewees back to the specifics of the critical incident improved as I proceeded through the 30 interviews. The same was true during the data analysis process—my ability to identify themes in the data, and to unify those themes, improved as the analysis progressed. However, by using grounded theory's constant comparative method to continually review and restructure themes, I was able to improve the analysis as my skills improved. This phenomenon of "researcher-as-instrument" is commonly accepted as part of qualitative research (Pezalla et al., 2012) and, accordingly, I continually monitored and reflected upon my own participation in the interviews and the data analysis.

8.3. Future work

8.3.1. Deep dives into key issues

Additional insights into each of the six key issues could be pursued in future research studies. To suggest just a few examples: Data quality was an issue that curators considered to be outside their purview; instead, they focused on metadata quality. However, research shows that curators may indeed support data quality and trustworthiness by facilitating standardized terminology, metadata, and formats, and by working with researchers to provide clear documentation of quality issues such as missing data, outliers, and inconsistencies. The issue of preserving the context of reused data is also one of the most complex and challenging issues addressed in this dissertation; this issue warrants additional research to develop strategies for preserving context in both qualitative and big social data. To enhance data comparability, more research and advocacy could be conducted to develop and operationalize interoperable, standardized metadata schemas.

8.3.2. Guidelines and policies for responsible big social research and qualitative data reuse

Our main ethical oversight mechanism for researchers in the United States is the IRB. However, IRBs are compliance bodies, not ethics boards; they can only help researchers comply with existing ethical standards. Unless those standards speak to big social research and qualitative data reuse, an IRB cannot provide the guidance needed for responsible research in these areas. Legislation and regulation may help, but the scholarly community needs to find ways to ensure epistemologically sound, ethical, and legal big social research and qualitative data reuse in the meantime. The fact that only two researchers interviewed for this dissertation referred to community ethics guidelines shows that such guidelines are not widely disseminated or adopted. Many professional organizations produce ethical guidelines, and the data curation community also produces guides such as the Data Curation Network data curation primers. However, these guidelines were rarely discussed by my interview participants, suggesting that these guidelines are not yet seen as standard practices to be adhered to. Future work for curators could include advocacy for standardized

data curation practices to support big social research and qualitative data reuse. Engaging with professional organizations such as Research Data Access and Preservation and the Digital Library Federation could support standardization in data curation practice. These practices could also be taught to the next generation of data curators through standardized curriculum in Library and Information Science graduate programs. As with any standard, the community will need to commit to regularly revising and updating these standard practices.

8.3.3. The changing social media landscape

Social media as a source of big social data is constantly changing. Users are now widely aware of the darker sides of social media, including data privacy issues, surveillance, dissemination of misinformation and disinformation, impact on elections, and the potential to cultivate violent fringe groups. Additionally, the popularity of social media platforms is constantly evolving: new types of platforms are emerging, and social media influencers have become a prominent user group in recent years. The conversation about Elon Musk's potential acquisition of Twitter in 2022 (Chotiner, 2022; Conger & Hirsch, 2022) highlights the commercial nature of social media platforms, and how a single wealthy buyer can change how a social media platform functions. Users may opt out of some platforms, user group demographics are changing, and the nature of user content is evolving. All of these factors will impact big social research. While big social data will continue to be available for the foreseeable future, researchers and curators will need to contend with a rapidly-evolving social media landscape, which will affect all six key issues addressed in this dissertation—context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. Future research could investigate how these six issues change depending on the social media landscape, or could suggest different issues depending on how social media changes. Data curation strategies may also need to be adjusted to support evolving data sources for big social research.

8.3.4. The value of small data

This dissertation operated under the assumption that scaling up research is an important goal. As Kitchin writes, qualitative data reuse and big social research both have the potential

to produce “studies with much greater breadth, depth, scale, timeliness and [which are] inherently longitudinal, in contrast to existing social sciences research” (Kitchin, 2014, p. 140). And as Housley et al. (2014) write, “The distinctive quality of big and broad social data for research is the possibilities it provides for the continuous (‘real-time’) observation of populations hitherto only accessible through episodic and retrospective snapshots gleaned through such instruments as household surveys and census data, longitudinal studies of cohorts and experiments measuring pre-test and post-test conditions” (p. 5).

However, Kitchin also writes that while “data infrastructures and big data will enhance the suite of data available for analysis and enable new approaches and techniques, [they] will not replace small data studies” (p. 148). Manovich (2012) also emphasizes that the depth of knowledge that can be gleaned from big data is not comparable to the depth that an ethnographer can plumb from embedding in a community. He concludes that big social data answers different questions from ethnographic or other in-depth social data. Housley et al. suggest that “the real transformative power of big and broad social data is in its use to augment and re-orientate rather than replace the other more established research strategies and designs” (2014, p. 5).

Scaling up research may not always be the ultimate goal. As boyd and Crawford write, “The size of data should fit the research question being asked; in some cases, small is best” (2012, p. 670)—an idea that applies to qualitative data reuse as well as big social research. In fact, scaling down big social datasets could alleviate some of the issues identified in this dissertation. For example, scaling down could enable informed consent for big social research, reduce the complexity of privacy and intellectual property issues, could allow for the collection of additional contextual information about social media users, and could increase data quality. More research could be done to consider how scale influences data curation for big social research and qualitative data reuse, and how data curators can engage with researchers to curate both big and small data.

8.4. Closing thoughts

As data sharing continues to grow, the key issues discussed in this dissertation will evolve in scope and complexity. Throughout this dissertation, I have focused on the data curators’ role

in supporting responsible research and data sharing. However, data curation practices can be adopted by anyone who is involved in the research process, and should be considered by all members of a research team. To help promote broad adoption of good data curation practices during a research project, research teams can engage (as an entire team) in data management planning prior to any data collection. Initiatives to embed curators into research projects and/or to designate specific research team members as data curation point-people can also support good data curation practices throughout the entire research lifecycle. That said, data curation is a growing profession, and an increasing number of trained data curators are well-positioned to lead data curation initiatives. The results of my research indicate that data curators should make additional effort to connect with researchers at every stage of the research lifecycle to encourage epistemologically sound, ethical, and legal big social research and qualitative data sharing and reuse. Data curators can speak about issues that matter to a variety of communities of practice, and thus begin to bridge gaps between these communities. Encouraging these connections between different communities of practice will lead to more responsible research, will increase data sharing and reuse, and will enhance discoveries in social and behavioral science.

References

Acker, A., & Kriesberg, A. (2017). Tweets may be archived: Civic engagement, digital preservation and Obama White House social media data. *Proceedings of the Association for Information Science and Technology*, *54*(1), 1–9. https://doi.org/10.1002/pra2.2017.14505401001

Akers, K. G., Read, K. B., Amos, L., Federer, L. M., Logan, A., & Plutchak, T. S. (2019). Announcing the Journal of the Medical Library Association's data sharing policy. *Journal of the Medical Library Association : JMLA*, *107*(4), 468–471. https://doi.org/10.5195/jmla.2019.801

Ako-Adjei, K., & Penna, M. (2021, October 5). *Microsoft Bookings*. https://web.archive.org/web/20211228185245/https://docs.microsoft.com/en-us/microsoft-365/bookings/bookings-overview?view=o365-worldwide

Alexander, S. M., Jones, K., Bennett, N. J., Budden, A., Cox, M., Crosas, M., Game, E. T., Geary, J., Hardy, R. D., Johnson, J. T., Karcher, S., Motzer, N., Pittman, J., Randell, H., Silva, J. A., da Silva, P. P., Strasser, C., Strawhacker, C., Stuhl, A., & Weber, N. (2020). Qualitative data sharing and synthesis for sustainability science. *Nature Sustainability*, *3*(2), 81–88. https://doi.org/10.1038/s41893-019-0434-8

Altman, I. (1977). Privacy regulation: Culturally universal or culturally specific? *Journal of Social Issues*, *33*(3), 66–84. https://doi.org/10.1111/j.1540-4560.1977.tb01883.x

Amer-Yahia, S., Doan, A., Kleinberg, J., Koudas, N., & Franklin, M. (2010). Crowds, clouds, and algorithms: Exploring the human side of "big data" applications. *Proceedings of the 2010 International Conference on Management of Data - SIGMOD '10*, 1259–1260. https://doi.org/10.1145/1807167.1807341

Andrejevic, M. (2014). The big data divide. *International Journal of Communication*, *8*(2014), 1673–1689. http://ijoc.org/index.php/ijoc/article/view/2161

APA Data Sharing Working Group. (2015). *Data sharing: Principles and considerations for policy development*. American Psychological Association. https://web.archive.org/web/20220120232933/https://www.apa.org/science/leadership/bsa/data-sharing-report

Ardichvili, A., Maurer, M., Li, W., Wentling, T., & Stuedemann, R. (2006). Cultural influences on knowledge sharing through online communities of practice. *Journal of Knowledge Management*, *10*(1), 94–107. https://doi.org/10.1108/13673270610650139

ASA. (2018). *American Sociological Association code of ethics*. American Sociological Association. https://web.archive.org/web/20220121025817/https://www.asanet.org/sites/defaul

t/files/asa_code_of_ethics-june2018.pdf

Ayres, L. (2008). Semi-structured interview. In L. Given (Ed.), *The SAGE encyclopedia of qualitative research methods* (pp. 810-811). SAGE Publications. https://doi.org/10.4135/9781412963909.n420

Baram-Tsabari, A., Segev, E., & Sharon, A. J. (2017). What's new? The applications of data mining and big data in the social sciences. In N. G. Fielding, R. M. Lee, & G. Blank (Eds.), *The SAGE handbook of online research methods* (pp. 92–106). SAGE Publications. https://doi.org/10.4135/9781473957992.n6

Barhorst, J. B., McLean, G., Brooks, J., & Wilson, A. (2019, June 5). Everyday micro-influencers and their impact on corporate brand reputation. *Proceedings of the 21st ICIG Symposium*. https://web.archive.org/web/20210422000726/https://strathprints.strath.ac.uk/68724/

Bates, M. J. (1999). The invisible substrate of information science. *Journal of the American Society for Information Science*, *50*(12), 1043–1050. https://doi.org/10.1002/(SICI)1097-4571(1999)50:12<1043::AID-ASI1>3.0.CO;2-X

Beagrie, N., & Houghton, J. (2014). The value and impact of data sharing and curation: A synthesis of three recent studies of UK research data centres. *Jisc Report*. https://web.archive.org/web/20220107080709/https://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf

Bechmann, A., & Lomborg, S. (2012). Mapping actor roles in social media: Different perspectives on value creation in theories of user participation: *New Media & Society 15*(5), 765-781. https://doi.org/10.1177/1461444812462853

Bechmann, A., & Vahlstrup, P. B. (2015). Studying Facebook and Instagram data: The Digital Footprints software. *First Monday*, *20*(12). https://doi.org/10.5210/fm.v20i12.5968

Beguerisse-Díaz, M., McLennan, A. K., Garduño-Hernández, G., Barahona, M., & Ulijaszek, S. J. (2017). The 'who' and 'what' of #diabetes on Twitter. *Digital Health*, *3*, 1-29. https://doi.org/10.1177/2055207616688841

Ben-David, A., & Huurdeman, H. (2014). Web archive search as research: Methodological and theoretical implications. *Alexandria*, *25*(1–2), 93–111. https://doi.org/10.7227/ALX.0022

Bernanke, B. S. (2004). Editorial statement. *The American Economic Review*, *94*(1), 404. https://www.jstor.org/stable/3592790

Bernard, H. R., Pelto, P. J., Werner, O., Boster, J., Romney, A. K., Johnson, A., Ember, C. R., & Kasakoff, A. (1986). The construction of primary data in cultural anthropology.

Current Anthropology, *27*(4), 382–396. https://doi.org/10.1086/203456

Bernard, H. R., Wutich, A., & Ryan, G. W. (2017). *Analyzing qualitative data: Systematic approaches* (2nd ed.). SAGE Publications.

Bill & Melinda Gates Foundation. (2015). *Open access policy*. Bill & Melinda Gates Foundation. https://web.archive.org/web/20220208135523/https://openaccess.gatesfoundation.org/open-access-policy/

Bishop, L. (2012). Using archived qualitative data for teaching: Practical and ethical considerations. *International Journal of Social Research Methodology*, *15*(4), 341–350. https://doi.org/10.1080/13645579.2012.688335

Bishop, L., & Kuula-Luumi, A. (2017). Revisiting qualitative data reuse: A decade on. *SAGE Open*, *7*(1), 1-15. https://doi.org/10.1177/2158244016685136

Blank, J. (2018, May 8). IP law in the age of social media. *Northeastern University Graduate Programs*. https://web.archive.org/web/20220123154530/https://www.northeastern.edu/graduate/blog/intellectual-property-and-social-media/

Bloomberg, L. D., & Volpe, M. (2016). *Completing your qualitative dissertation: A road map from beginning to end* (3rd ed.). SAGE Publications.

BMJ. (1996). The Nuremberg Code (1947). *BMJ*, *313*(7070), 1448–1448. https://doi.org/10.1136/bmj.313.7070.1448

Bogen, K. W., Bleiweiss, K. K., Leach, N. R., & Orchowski, L. M. (2021). #MeToo: Disclosure and response to sexual victimization on Twitter. *Journal of Interpersonal Violence*, *36*(17–18), 8257–8288. https://doi.org/10.1177/0886260519851211

Bogner, A., Littig, B., & Menz, W. (2009). Introduction: Expert interviews — An introduction to a new methodological debate. In A. Bogner, B. Littig, & W. Menz (Eds.), *Interviewing experts* (pp. 1–13). Palgrave Macmillan UK. https://doi.org/10.1057/9780230244276_1

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295–298. https://doi.org/10.1038/nature11421

Borgen, W. A., Amundson, N. E., & Butterfield, L. D. (2008). Critical incident technique. In L. Given, *The SAGE encyclopedia of qualitative research methods* (pp. 158-159). SAGE Publications. https://doi.org/10.4135/9781412963909.n84

Borges-Rey, E. (2016). Unravelling data journalism. *Journalism Practice*, *10*(7), 833–843.

https://doi.org/10.1080/17512786.2016.1159921

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. https://doi.org/10.1002/asi.22634

Bos, N., Zimmerman, A., Olson, J., Yew, J., Yerkie, J., Dahl, E., & Olson, G. (2007). From shared databases to communities of practice: A taxonomy of collaboratories. *Journal of Computer-Mediated Communication*, *12*(2), 652–672. https://doi.org/10.1111/j.1083-6101.2007.00343.x

Bosher, H., & Yeşiloğlu, S. (2019). An analysis of the fundamental tensions between copyright and social media: The legal implications of sharing images on Instagram. *International Review of Law, Computers & Technology*, *33*(2), 164–186. https://doi.org/10.1080/13600869.2018.1475897

Bossetta, M. (2018). The digital architectures of social media: Comparing political campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. election. *Journalism & Mass Communication Quarterly*, *95*(2), 471–496. https://doi.org/10.1177/1077699018763307

Bourdieu, P. (1986). The forms of capital. In J. Richardson (Ed.), *Handbook of theory and research for the sociology of education* (pp. 241–258). Greenwood.

boyd, d. (2013). *Bibliography of research on Twitter & microblogging*. https://web.archive.org/web/20191123145930/https://www.danah.org/researchBibs/twitter.php

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

boyd, d., & Ellison, N. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

Bright, J. (2017). 'Big social science': Doing big data in the social sciences. In N. G. Fielding, R. M. Lee, & G. Blank (Eds.), *The SAGE handbook of online research methods* (pp. 125–139). SAGE Publications. https://doi.org/10.4135/9781473957992.n8

Broom, A., Cheshire, L., & Emmison, M. (2009). Qualitative researchers' understandings of their practice and the implications for data archiving and sharing. *Sociology*, *43*(6), 1163–1180. https://doi.org/10.1177/0038038509345704

Bruns, A. (2013). Faster than the speed of print: Reconciling 'big data' social media analysis and academic scholarship. *First Monday*, *18*(10).

https://doi.org/10.5210/fm.v18i10.4879

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Bruns, A., & Weller, K. (2016). Twitter as a first draft of the present: And the challenges of preserving it for the future. In S. Staab & P. Parigi (Eds.), *Proceedings of the 8th ACM Conference on Web Science* (pp. 183–189). https://doi.org/10.1145/2908131

Buchanan, E. (2017). Internet research ethics: Twenty years later. In M. Zimmer & K. Kinder-Kurlanda (Eds.), *Internet research ethics for the social age: New challenges, cases, and contexts* (pp. ix–xv). Peter Lang.

Burgess, J., & Bruns, A. (2012). Twitter archives and the challenges of "big social data" for media and communication research. *M/C Journal*, *15*(5), Article 5. https://doi.org/10.5204/mcj.561

Butterfield, L. D., Borgen, W. A., Amundson, N. E., & Maglio, A.-S. T. (2005). Fifty years of the critical incident technique: 1954-2004 and beyond. *Qualitative Research*, *5*(4), 475–497. https://doi.org/10.1177/1468794105056924

Buysse, V., Sparkman, K. L., & Wesley, P. W. (2003). Communities of practice: Connecting what we know with what we do. *Exceptional Children*, *69*(3), 263–277. https://doi.org/10.1177/001440290306900301

Caliandro, A. (2018). Digital methods for ethnography: Analytical concepts for ethnographers exploring social media environments. *Journal of Contemporary Ethnography*, *47*(5), 551–578. https://doi.org/10.1177/0891241617702960

Cappella, J. N. (2017). Vectors into the future of mass and interpersonal communication research: Big data, social media, and computational social science. *Human Communication Research*, *43*(4), 545–558. https://doi.org/10.1111/hcre.12114

Carroll, S. R., Herczog, E., Hudson, M., Russell, K., & Stall, S. (2021). Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Scientific Data*, *8*(1), 108. https://doi.org/10.1038/s41597-021-00892-0

Castells, M. (2000). Materials for an exploratory theory of the network society. *The British Journal of Sociology*, *51*(1), 5–24. https://doi.org/10.1111/j.1468-4446.2000.00005.x

Castillo, D., Coates, H., & Narlock, M. (2021). Qualitative data curation primer. *Data Curation Network*. https://hdl.handle.net/11299/219053

Cavazos-Rehg, P. A., Krauss, M., Fisher, S. L., Salyer, P., Grucza, R. A., & Bierut, L. J. (2015). Twitter chatter about marijuana. *Journal of Adolescent Health*, *56*(2), 139–145.

https://doi.org/10.1016/j.jadohealth.2014.10.270

Center for Qualitative and Multi-Method Inquiry. (2020). *Qualitative Data Repository*. https://web.archive.org/web/20220427033603/https://qdr.syr.edu/

CERN Data Centre. (2020). *Zenodo*. https://web.archive.org/web/20200524175824/https://zenodo.org/

Chang, Y.-W. (2018). Exploring the interdisciplinary characteristics of library and information science (LIS) from the perspective of interdisciplinary LIS authors. *Library & Information Science Research*, *40*(2), 125–134. https://doi.org/10.1016/j.lisr.2018.06.004

Charmaz, K. (2001). Qualitative interviewing and grounded theory analysis. In J. Gubrium & J. Holstein, *Handbook of interview research* (pp. 675–694). SAGE Publications. https://doi.org/10.4135/9781412973588.n39

Charmaz, K. (2008). Reconstructing grounded theory. In P. Alasuutari, L. Bickman, & J. Brannen, *The SAGE handbook of social research methods* (pp. 461–478). SAGE Publications. https://doi.org/10.4135/9781446212165.n27

Chawinga, W. D., & Zinn, S. (2019). Global perspectives of research data sharing: A systematic literature review. *Library & Information Science Research*, *41*(2), 109–122. https://doi.org/10.1016/j.lisr.2019.04.004

Chawla, N. V., & Davis, D. A. (2013). Bringing big data to personalized healthcare: A patient-centered framework. *Journal of General Internal Medicine*, *28*(3), 660–665. https://doi.org/10.1007/s11606-013-2455-8

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, *36*(4), 1165–1188. https://doi.org/10.2307/41703503

Chin, G., Jr., & Lansing, C. S. (2004). Capturing and supporting contexts for scientific data sharing via the biological sciences collaboratory. *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, 409–418. https://doi.org/10.1145/1031607.1031677

Chotiner, I. (2022, April 26). Why Elon Musk bought Twitter. *The New Yorker*. https://web.archive.org/web/20220505160034/https://www.newyorker.com/news/q-and-a/why-elon-musk-bought-twitter

Chu, K.-H., Colditz, J., Sidani, J., Zimmer, M., & Primack, B. (2021). Re-evaluating standards of human subjects protection for sensitive health data in social media networks. *Social Networks 67*(October 2021), 41-46. https://doi.org/10.1016/j.socnet.2019.10.010

Clark, A. (2006). Anonymising research data (NCRM working paper series). *ESRC National Centre for Research Methods*. https://eprints.ncrm.ac.uk/id/eprint/480

Clark, K., Duckham, M., Guillemin, M., Hunter, A., McVernon, J., O'Keefe, C., Pitkin, C., Prawer, S., Sinnott, R., Warr, D., & Waycott, J. (2019). Advancing the ethical use of digital data in human research: Challenges and strategies to promote ethical practice. *Ethics and Information Technology 21*, 59–73. https://doi.org/10.1007/s10676-018-9490-4

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, *64*(2), 317–332. https://doi.org/10.1111/jcom.12084

Colombo, G. B., Burnap, P., Hodorog, A., & Scourfield, J. (2016). Analysing the connectivity and communication of suicidal users on Twitter. *Computer Communications*, *73*, 291–300. https://doi.org/10.1016/j.comcom.2015.07.018

Conger, K., & Hirsch, L. (2022, July 12). Twitter Sues Musk After He Tries Backing Out of $44 Billion Deal. *The New York Times.* https://web.archive.org/web/20220818103701/https://www.nytimes.com/2022/07/12/technology/twitter-lawsuit-musk-acquisition.html

Cooky, C., Linabary, J. R., & Corple, D. J. (2018). Navigating big data dilemmas: Feminist holistic reflexivity in social media research. *Big Data & Society*, *5*(2), 1-12. https://doi.org/10.1177/2053951718807731

Cooper, C., Booth, A., Britten, N., & Garside, R. (2017). A comparison of results of empirical studies of supplementary search techniques and recommendations in review methodology handbooks: A methodological review. *Systematic Reviews*, *6*(1), Article 234. https://doi.org/10.1186/s13643-017-0625-1

Cooper, H. M., Hedges, L. V., & Valentine, J. C. (Eds.). (2019). *Handbook of research synthesis and meta-analysis* (3rd ed.). Russell Sage Foundation.

Corbin, J., & Strauss, A. (2008). Theoretical sampling. In *Basics of qualitative research: Techniques and procedures for developing grounded theory* (3rd ed., pp. 133-158). SAGE Publications. https://doi.org/10.4135/9781452230153.n7

CoreTrustSeal. (2020). *Core trustworthy data repositories requirements*. https://web.archive.org/web/20200408004456/https://www.coretrustseal.org/why-certification/requirements/

Cornell Research Services. (2019). *IRB consent form templates*. https://web.archive.org/web/20210609195943/https://researchservices.cornell.edu/forms/irb-consent-form-templates

Corral, M. (2020). Atlas.ti data curation primer. *Data Curation Network*. https://hdl.handle.net/11299/210211

Corti, L. (1999). Text, sound and videotape: The future of qualitative data in the global network. *IASSIST Quarterly*, *23*(2), 15. https://doi.org/10.29173/iq726

Corti, L. (2000). Progress and problems of preserving and providing access to qualitative data for social research—The international picture of an emerging culture. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *1*(3), Article 3. https://doi.org/10.17169/fqs-1.3.1019

Corti, L., & Backhouse, G. (2005). Acquiring qualitative data for secondary analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *6*(2), Article 2. https://doi.org/10.17169/fqs-6.2.459

Corti, L., & Gregory, A. (2011). CAQDAS comparability. What about CAQDAS data exchange? *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *12*(1). https://doi.org/10.17169/FQS-12.1.1634

Corti, L., & Thompson, P. (1996). ESRC Qualitative Data Archival Resource Centre (QUALIDATA). *Sociological Research Online*. https://web.archive.org/web/20210509204938/https://www.socresonline.org.uk/1/3/qualidata.html

Corti, L., & Thompson, P. (1998). Are you sitting on your qualitative data? Qualidata's mission. *International Journal of Social Research Methodology*, *1*(1), 85–89. https://doi.org/10.1080/13645579.1998.10846865

Corti, L., Witzel, A., & Bishop, L. (2005). On the potentials and problems of secondary analysis. An introduction to the FQS special issue on secondary analysis of qualitative data. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *6*(1), Article 1. https://doi.org/10.17169/fqs-6.1.498

Cox, A. M., Kennan, M. A., Lyon, L., & Pinfield, S. (2017). Developments in research data management in academic libraries: Towards an understanding of research data service maturity. *Journal of the Association for Information Science and Technology*, *68*(9), 2182–2200. https://doi.org/10.1002/asi.23781

Creative Commons. (2014). *CC0 use for data*. https://web.archive.org/web/20220424152506/https://wiki.creativecommons.org/wiki/CC0_use_for_data

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed). SAGE Publications.

Croeser, S., & Highfield, T. (2020). Blended data: Critiquing and complementing social media

datasets, big and small. In J. Hunsinger, M. M. Allen, & L. Klastrup (Eds.), *Second international handbook of internet research* (pp. 669–690). Springer Netherlands. https://doi.org/10.1007/978-94-024-1555-1_15

Cronin, B. (2008). The sociological turn in information science. *Journal of Information Science*, *34*(4), 465–475. https://doi.org/10.1177/0165551508088944

Cushing, A. L., & Dumbleton, O. (2017). 'We have to make an effort with it': Exploring the use of stages to help understand the personal information management needs of humanities and social science doctoral students managing dissertation information. *IFLA Journal*, *43*(1), 40–50. https://doi.org/10.1177/0340035216686983

Dale, A., Arber, S., & Procter, M. (1988). *Doing secondary analysis*. Allen & Unwin.

Darragh, J., Hofelich Mohr, A., Hunt, S., Woodbrook, R., Fearon, D., Moore, J., & Hadley, H. (2020). Human subjects data essentials data curation primer. *Data Curation Network*. https://hdl.handle.net/11299/216579

Davidson, E., Edwards, R., Jamieson, L., & Weller, S. (2018). Big data, qualitative style: A breadth-and-depth method for working with large amounts of secondary qualitative data. *Quality & Quantity*, *53*, 363–376. https://doi.org/10.1007/s11135-018-0757-y

DDI Alliance. (2019). *Data Documentation Initiative*. https://web.archive.org/web/20220202185335/https://ddialliance.org/

de Lusignan, S., Chan, T., Theadom, A., & Dhoul, N. (2007). The roles of policy and professionalism in the protection of processed clinical data: A literature review. *International Journal of Medical Informatics*, *76*(4), 261–268. https://doi.org/10.1016/j.ijmedinf.2005.11.003

Demgenski, R., Karcher, S., Kirilova, D., & Weber, N. (2021). Introducing the Qualitative Data Repository's curation handbook. *Journal of EScience Librarianship*, *10*(3), e1207. https://doi.org/10.7191/jeslib.2021.1207

di Gregorio, S. (2019, March 25). *Unlocking the power of qualitative data for future generations*. QSR International, NVivo Blog. https://web.archive.org/web/20220510153708/https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/resources/blog/unlocking-the-power-of-qualitative-data-for-future

Diebold, F. X. (2012). A personal perspective on the origin(s) and development of "big data": The phenomenon, the term, and the discipline, second version. *PIER Working Paper No. 13-003*. https://doi.org/10.2139/ssrn.2202843

DocNow. (2020). *Documenting the Now*. https://web.archive.org/web/20220419155938/https://www.docnow.io/

DocuSign. (2021). *Docusign product features*. https://web.archive.org/web/20211008211934/https://www.docusign.com/features-and-benefits/features

Doft, D. (2015). Facebook, Twitter, and the Wild West of IP enforcement on social media: Weighing the merits of a uniform dispute resolution policy. 49 *John Marshall Law Review* 959. https://repository.law.uic.edu/lawreview/vol49/iss4/2/

Drakonakis, K., Ilia, P., Ioannidis, S., & Polakis, J. (2019). Please forget where I was last summer: The privacy risks of public location (meta)data. *Proceedings of the 2019 Network and Distributed System Security Symposium,* 1-15. https://doi.org/10.14722/ndss.2019.23151

Driscoll, K., & Walker, S. (2014). Working within a black box: Transparency in the collection and production of big Twitter data. *International Journal of Communication*, *8*(2014), 1745–1764. https://ijoc.org/index.php/ijoc/article/view/2171

Dryad. (2022). *Dryad Digital Repository*. https://web.archive.org/web/20200524165914/https://datadryad.org/stash

Dryad Digital Repository. (2011). *Joint Data Archiving Policy*. http://wiki.datadryad.org/Joint_Data_Archiving_Policy_(JDAP)

DuBois, J. M., Strait, M., & Walsh, H. (2018). Is it time to share qualitative research data? *Qualitative Psychology*, *5*(3), 380–393. https://doi.org/10.1037/qup0000076

Duke, C. S., & Porter, J. H. (2013). The ethics of data sharing and reuse in biology. *BioScience*, *63*(6), 483–489. https://doi.org/10.1525/bio.2013.63.6.10

Dunn, C. S., & Austin, E. W. (1998). Protecting confidentiality in archival data resources. *IASSIST Quarterly*, *22*(2), 16-24. https://doi.org/10.29173/iq724

Durdella, N. (2019). Developing data collection instruments and describing data collection procedures. In *Qualitative dissertation methodology: A guide for research design and methods* (pp. 1–43). SAGE Publications. https://doi.org/10.4135/9781506345147

Ellard-Gray, A., Jeffrey, N. K., Choubak, M., & Crann, S. E. (2015). Finding the hidden participant: Solutions for recruiting hidden, hard-to-reach, and vulnerable populations. *International Journal of Qualitative Methods*, *14*(5), 1–10. https://doi.org/10.1177/1609406915621420

Ellison, N., Heino, R., & Gibbs, J. (2006). Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer-Mediated Communication*, *11*(2), 415–441. https://doi.org/10.1111/j.1083-6101.2006.00020.x

Elman, C., Hoelter, L., Kapiszewski, D., & Kirilova, D. (2017, November 5). IRB guidelines and

data sharing in the social science: Tensions and strategies to address them. PRIM&R Social, Behavioral, and Educational Research Conference. https://doi.org/10.6084/m9.figshare.5969104.v1

Elman, C., & Kapiszewski, D. (2014). Data access and research transparency in the qualitative tradition. *PS: Political Science & Politics*, *47*(1), 43–47. https://doi.org/10.1017/S1049096513001777

Elman, C., Kapiszewski, D., & Lupia, A. (2018). Transparent social inquiry: Implications for political science. *Annual Review of Political Science*, *21*(1), 29–47. https://doi.org/10.1146/annurev-polisci-091515-025429

Elman, C., Kapiszewski, D., & Vinuela, L. (2010). Qualitative data archiving: Rewards and challenges. *PS: Political Science & Politics*, *43*(1), 23–27. https://doi.org/10.1017/S104909651099077X

Ember, C. R. (2007). Using the HRAF collection of ethnography in conjunction with the Standard Cross-Cultural Sample and the Ethnographic Atlas. *Cross-Cultural Research*, *41*(4), 396–427. https://doi.org/10.1177/1069397107306593

Evers, J. C. (2018). Current issues in qualitative data analysis software (QDAS): A user and developer perspective. *The Qualitative Report*, *23*(13), 61–73. https://doi.org/10.46743/2160-3715/2018.3205

Evers, J., Caprioli, M. U., Nöst, S., & Wiedemann, G. (2020). What is the REFI-QDA standard: Experimenting with the transfer of analyzed research projects between QDA software. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *21*(2). https://doi.org/10.17169/FQS-21.2.3439

Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM*, *57*(6), 74–81. https://doi.org/10.1145/2602574

Faniel, I. M., & Connaway, L. (2018). Librarians' perspectives on the factors influencing research data management programs. *College & Research Libraries*, *79*(1), 100-119. https://doi.org/10.5860/crl.79.1.100

Faniel, I. M., Frank, R. D., & Yakel, E. (2019). Context from the data reuser's point of view. *Journal of Documentation, 75*(6), 1274-1297. https://doi.org/10.1108/JD-08-2018-0133

Faniel, I. M., Kriesberg, A., & Yakel, E. (2016). Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science and Technology*, *67*(6), 1404–1416. https://doi.org/10.1002/asi.23480

Fielding, N. (2004). Getting the most from archived qualitative data: Epistemological, practical and professional obstacles. *International Journal of Social Research*

Methodology, *7*(1), 97–104. https://doi.org/10.1080/13645570310001640699

Fielding, N., & Fielding, J. L. (2000). Resistance and adaptation to criminal identity: Using secondary analysis to evaluate classic studies of crime and deviance. *Sociology*, *34*(4), 671–689. https://doi.org/10.1177/S0038038500000419

Fienberg, S. E., Martin, M. E., & Straf, M. L. (Eds.). (1985). *Sharing research data*. The National Academies Press. https://doi.org/10.17226/2033

Fiesler, C., Dye, M., Feuston, J. L., Hiruncharoenvate, C., Hutto, C. J., Morrison, S., Khanipour Roshan, P., Pavalanathan, U., Bruckman, A. S., De Choudhury, M., & Gilbert, E. (2017). What (or who) is public?: Privacy settings and social media content sharing. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 567–580. https://doi.org/10.1145/2998181.2998223

Fiesler, C., & Proferes, N. (2018). "Participant" perceptions of Twitter research ethics. *Social Media + Society*, *4*(1), 1-14. https://doi.org/10.1177/2056305118763366

Fink, A. S. (2000). The role of the researcher in the qualitative research process. A potential barrier to archiving qualitative data. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *1*(3), Article 3. https://doi.org/10.17169/fqs-1.3.1021

FNIGC. (2010). *The First Nations Principles of OCAP®, a registered trademark of the First Nations Information Governance Centre (FNIGC)*. First Nations Information Governance Centre. https://web.archive.org/web/20220312085249/https://fnigc.ca/ocap-training/

Frank, R. D., Chen, Z., Crawford, E., Suzuka, K., & Yakel, E. (2017). Trust in qualitative data repositories. *Proceedings of the Association for Information Science and Technology*, *54*(1), 102–111. https://doi.org/10.1002/pra2.2017.14505401012

Franzke, A. S., Bechmann, A., Ess, C. M., & Zimmer, M. (2020). Internet research: Ethical guidelines 3.0. *The International Association of Internet Researchers (AoIR)*. https://web.archive.org/web/20220402125705/https://aoir.org/reports/ethics3.pdf

Fuchs, C. (2017). *Social media: A critical introduction*. SAGE Publications.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Garfinkel, H. (1967). *Studies in ethnomethodology*. Polity Press.

Garfinkel, S. L. (2015). *De-identification of personal information* (NIST IR 8053). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8053

Ghermandi, A., & Sinclair, M. (2019). Passive crowdsourcing of social media in environmental research: A systematic map. *Global Environmental Change*, *55*, 36–47. https://doi.org/10.1016/j.gloenvcha.2019.02.003

Giarlo, M. (2013). Academic libraries as data quality hubs. *Journal of Librarianship and Scholarly Communication*, *1*(3), eP1059. https://doi.org/10.7710/2162-3309.1059

Glaser, B. G. (1962). Secondary analysis: A strategy for the use of knowledge from research elsewhere. *Social Problems*, *10*(1), 70–74. https://doi.org/10.2307/799409

Glaser, B. G. (1963). Retreading research materials: The use of secondary analysis by the independent researcher. *American Behavioral Scientist*, *6*(10), 11–14. https://doi.org/10.1177/000276426300601003

Glaser, B. G., & Strauss, A. L. (1967). *Discovery of grounded theory: Strategies for qualitative research*. Aldine.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–8. https://doi.org/10.2307/1174772

Glenna, L., Hesse, A., Hinrichs, C., Chiles, R., & Sachs, C. (2019). Qualitative research ethics in the big data era. *American Behavioral Scientist*, *63*(5), 555–559. https://doi.org/10.1177/0002764219826282

Golder, S., Scantlebury, A., & Christmas, H. (2019). Understanding public attitudes toward researchers using social media for detecting and monitoring adverse events data: Multi methods study. *Journal of Medical Internet Research*, *21*(8), e7081. https://doi.org/10.2196/jmir.7081

González-Bailón, S. (2013). Social science in the era of big data. *Policy & Internet*, *5*(2), 147–160. https://doi.org/10.1002/1944-2866.POI328

Goodwin, J. (2012). *SAGE secondary data analysis*. SAGE Publications.

Goodwin, J., & O'Connor, H. (2006). Contextualising the research process: Using interviewer notes in the secondary analysis of qualitative data. *The Qualitative Report, 11*(2), 374-392. https://doi.org/10.46743/2160-3715/2006.1679

Gray, J., Bounegru, L., & Chambers, L. (Eds.). (2012). *The data journalism handbook*. European Journalism Centre. https://web.archive.org/web/20210827005652/https://s3.eu-central-1.amazonaws.com/datajournalismcom/handbooks/The-Data-Journalism-Handbook-1.pdf

Greene, T., Shmueli, G., Ray, S., & Fell, J. (2019). Adjusting to the GDPR: The impact on data scientists and behavioral researchers. *Big Data*, *7*(3), 140–162. https://doi.org/10.1089/big.2018.0176

Greener, I. (2011). *Designing social research: A guide for the bewildered*. SAGE Publications. https://doi.org/10.4135/9781446287934

Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communication and Technology*, *29*(2), 75–91. JSTOR. https://www.jstor.org/stable/30219811

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. SAGE Publications.

Hadley, H. (2020). NVivo data curation primer. *Data Curation Network*. https://hdl.handle.net/11299/216583

Hakim, C. (1982). *Secondary analysis in social research: A guide to data sources and methods with examples*. Allen & Unwin.

Halavais, A. (2015). Bigger sociological imaginations: Framing big social data theory and methods. *Information, Communication & Society*, *18*(5), 583–594. https://doi.org/10.1080/1369118X.2015.1008543

Halford, S., & Savage, M. (2017). Speaking sociologically with big data: Symphonic social science and the future for big data research. *Sociology*, *51*(6), 1132–1148. https://doi.org/10.1177/0038038517698639

Hammersley, M. (1997). Qualitative data archiving: Some reflections on its prospects and problems. *Sociology*, *31*(1), 131–142. https://doi.org/10.1177/0038038597031001010

Hammersley, M. (2010). Can we re-use qualitative data via secondary analysis? Notes on some terminological and substantive issues. *Sociological Research Online*, *15*(1), 1–7. https://doi.org/10.5153/sro.2076

Hargittai, E. (2020). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review*, *38*(1), 10–24. https://doi.org/10.1177/0894439318788322

Hartter, J., Ryan, S. J., MacKenzie, C. A., Parker, J. N., & Strasser, C. A. (2013). Spatially explicit data: Stewardship and ethical challenges in science. *PLOS Biology*, *11*(9), e1001634. https://doi.org/10.1371/journal.pbio.1001634

Heaton, J. (1998). Secondary analysis of qualitative data. *Social Research Update*, *22*. https://web.archive.org/web/20220128133742/https://sru.soc.surrey.ac.uk/SRU22.html

Heaton, J. (2000, August). Secondary analysis of qualitative data: A review of the literature. *Social Policy Research Unit, University of York*, ESRC 1752 JH 8.00, REF: R000222918.

Heaton, J. (2004). *Reworking qualitative data*. SAGE Publications.

https://doi.org/10.4135/9781849209878

Heaton, J. (2008). Secondary analysis of qualitative data: An overview. *Historical Social Research / Historische Sozialforschung*, *33*(3(125)), 33–45. https://www.jstor.org/stable/20762299

Heidler, R. (2017). Epistemic cultures in conflict: The case of astronomy and high energy physics. *Minerva*, *55*(3), 249–277. https://doi.org/10.1007/s11024-017-9315-3

HELPS Lab. (2020). *The Human Ecology Learning and Problem Solving (HELPS) Lab at Montana State University—About Us*. https://web.archive.org/web/20200920231427/https://helpslab.montana.edu/About_Us.html

Hemphill, L., Leonard, S. H., & Hedstrom, M. (2018, June). Developing a social media archive at ICPSR. *Proceedings of Web Archiving and Digital Libraries (WADL'18)*. https://hdl.handle.net/2027.42/143185

Hewson, C., Vogel, C., & Laurent, D. (2016). Internet-mediated research: State of the art. In *Internet research methods* (pp. 33-70). SAGE Publications. https://doi.org/10.4135/9781473920804.n3

Hinds, P. S., Chaves, D. E., & Cypess, S. M. (1992). Context as a source of meaning and understanding. *Qualitative Health Research*, *2*(1), 61–74. https://doi.org/10.1177/104973239200200105

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. https://doi.org/10.1177/104973239700700306

hiQ Labs, Inc v. LinkedIn Corporation, 938 Federal Reporter 3rd 985 (9th Circuit 2019). https://web.archive.org/web/20220418134148/https://cdn.ca9.uscourts.gov/datastore/opinions/2019/09/09/17-16783.pdf

Hirtle, P. B., Hudson, E., & Kenyon, A. T. (2009). *Copyright and cultural institutions: Guidelines for digitization for U.S. libraries, archives, and museums*. Cornell University Library. https://ecommons.cornell.edu/handle/1813/14142

Hogan, B. (2010). The presentation of self in the age of social media: Distinguishing performances and exhibitions online. *Bulletin of Science, Technology & Society*, *30*(6), 377–386. https://doi.org/10.1177/0270467610385893

Holdren, J. P. (2013). Increasing access to the results of federally funded scientific research. *White House Office of Science and Technology Policy*. https://web.archive.org/web/20220427151449/https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research

Holland, J., Thomson, R., Henderson, S., London South Bank University, & Families & Social Capital ESRC Research Group. (2006). *Qualitative longitudinal research: A discussion paper*. London South Bank University.

Housley, W., Procter, R., Edwards, A., Burnap, P., Williams, M., Sloan, L., Rana, O., Morgan, J., Voss, A., & Greenhill, A. (2014). Big and broad social data and the sociological imagination: A collaborative response. *Big Data & Society*, *1*(2), 1-15. https://doi.org/10.1177/2053951714545135

Hu, X., Rousseau, R., & Chen, J. (2011). On the definition of forward and backward citation generations. *Journal of Informetrics*, *5*(1), 27–36. https://doi.org/10.1016/j.joi.2010.07.004

Hughes, H., Williamson, K., & Lloyd, A. (2007). Critical incident technique. In S. Lipu (Ed.), *Exploring methods in information literacy research* (pp. 49–66). Chandos Publishing.

Hutton, L., & Henderson, T. (2013). An architecture for ethical and privacy-sensitive social network experiments. *SIGMETRICS Performance Evaluation Review*, *40*(4), 90–95. https://doi.org/10.1145/2479942.2479954

Hyman, H. H. (1972). *Secondary analysis of sample surveys: Principles, procedures, and potentialities*. Wiley.

ICPSR. (2012). *Guide to social science data preparation and archiving: Introduction*. https://web.archive.org/web/20220121040859/https://www.icpsr.umich.edu/files/deposit/dataprep.pdf

ICPSR. (2019). *ICPSR: A case study in repository management*. https://web.archive.org/web/20190615220105/https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/index.html

ICPSR. (2022). *ICPSR, part of the Institute for Social Research at the University of Michigan*. https://web.archive.org/web/20220409021118/https://www.icpsr.umich.edu/web/pages/

IDS Association. (2022). *International Data Spaces: The future of the data economy is here*. https://web.archive.org/web/20220414092731/https://internationaldataspaces.org/

Ip, R. K. F., & Wagner, C. (2008). Weblogging: A study of social computing and its impact on organizations. *Decision Support Systems*, *45*, 242–250. https://doi.org/10.1016/j.dss.2007.02.004

Irwin, S. (2013). Qualitative secondary data analysis: Ethics, epistemology and context. *Progress in Development Studies*, *13*(4), 295–306. https://doi.org/10.1177/1464993413490479

Irwin, S., & Winterton, M. (2011). Debates in qualitative secondary analysis: Critical reflections. *Timescapes Working Paper Series*, *4*. https://doi.org/10.5518/200/04

Ishikawa, H. (2015). *Social big data mining*. CRC Press.

Ito, M. (2008). Introduction. In K. Varnelis (Ed.), *Networked publics* (pp. 1–14). MIT Press.

Jendryke, M., Balz, T., McClure, S. C., & Liao, M. (2017). Putting people in the picture: Combining big location-based social media data and remote sensing imagery for enhanced contextual urban information in Shanghai. *Computers, Environment and Urban Systems*, *62*, 99–112. https://doi.org/10.1016/j.compenvurbsys.2016.10.004

Johnston, L. R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R., Stewart, C., Blake, M., Herndon, J., McGeary, T. M., & Hull, E. (2018). Data Curation Network: A cross-institutional staffing model for curating research data. *International Journal of Digital Curation*, *13*, 125–140. https://doi.org/10.2218/ijdc.v13i1.616

Joly, Y., Dalpé, G., So, D., & Birko, S. (2015). Fair shares and sharing fairly: A survey of public views on open science, informed consent and participatory research in biobanking. *PLOS ONE*, *10*(7), e0129893. https://doi.org/10.1371/journal.pone.0129893

Jones, K., Alexander, S. M., Bennett, N., Bishop, L., Budden, A., Cox, M., Crosas, M., Game, E., Geary, J., Hahn, C., Hardy, D., Johnson, J., Karcher, S., LaFevor, M., Motzer, N., Pinto da Silva, P., Pittman, J., Randell, H., Silva, J., … Winslow, D. (2018). Qualitative data sharing and re-use for socio-environmental systems research: A synthesis of opportunities, challenges, resources and approaches. *SESYNC White Paper.* https://doi.org/10.13016/M2WH2DG59

Jules, B., Summers, E., & Mitchell, V. Jr. (2018). Ethical considerations for archiving social media content generated by contemporary social movements: Challenges, opportunities, and recommendations. *Documenting the Now White Paper*. https://web.archive.org/web/20220316220447/https://www.docnow.io/docs/docnow-whitepaper-2018.pdf

Kadushin, C. (1968). Power, influence and social circles: A new methodology for studying opinion makers. *American Sociological Review*, *33*(5), 685–699. https://doi.org/10.2307/2092880

Kansa, S. W., & Kansa, E. C. (2018). Data beyond the archive in digital archaeology: An introduction to the special section. *Advances in Archaeological Practice*, *6*(2), 89–92. https://doi.org/10.1017/aap.2018.7

Karcher, S., Kirilova, D., & Weber, N. (2016). Beyond the matrix: Repository services for qualitative data. *IFLA Journal*, *42*(4), 292–302. https://doi.org/10.1177/0340035216672870

Karcher, S., & Weber, N. (2019). Annotation for transparent inquiry: Transparent data and analysis for qualitative research. *IASSIST Quarterly*, *43*(2), 1–9. https://doi.org/10.29173/iq959

Keller, R., & Poferl, A. (2016). Epistemic cultures in sociology between individual inspiration and legitimization by procedure: Developments of qualitative and interpretive research in German and French sociology since the 1960s. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *17*(1), Article 1. https://doi.org/10.17169/fqs-17.1.2419

Kiecolt, J. K., & Nathan, L. E. (1985). *Secondary analysis of survey data*. SAGE Publications.

Kim, W., Jeong, O.-R., & Lee, S.-W. (2010). On social web sites. *Information Systems*, *35*(2), 215–236. https://doi.org/10.1016/j.is.2009.08.003

Kinder-Kurlanda, K., Weller, K., Zenk-Möltgen, W., Pfeffer, J., & Morstatter, F. (2017). Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. *Big Data & Society*, *4*(2), 1-14. https://doi.org/10.1177/2053951717736336

King, G., & Persily, N. (2020). A new model for industry–academic partnerships. *PS: Political Science & Politics*, *53*(4), 703-709. https://doi.org/10.1017/S1049096519001021

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures & their consequences*. SAGE Publications.

Knorr-Cetina, K. (1999). *Epistemic cultures: How the sciences make knowledge*. Harvard University Press.

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Lanchester, J. (2017, August 16). You are the product. *London Review of Books*, *39*(16). https://web.archive.org/web/20220407085326/https://www.lrb.co.uk/the-paper/v39/n16/john-lanchester/you-are-the-product

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *Meta Group*. https://web.archive.org/web/20190319224043/https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Latour, B. (1996). On actor-network theory: A few clarifications. *Soziale Welt*, *47*(4), 369–381. https://www.jstor.org/stable/40878163

Latour, B. (2007). Beware, your imagination leaves digital traces. *Times Higher Literary Supplement*, *6*(4), 129–131.

Lave, J., & Wenger, E. (1991). *Situated Learning Legitimate Peripheral Participation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815355

Lavori, P. W., Sugarman, J., Hays, M. T., & Feussner, J. R. (1999). Improving informed consent in clinical trials: A duty to experiment. *Controlled Clinical Trials*, *20*(2), 187–193. https://doi.org/10.1016/S0197-2456(98)00064-6

Lawton, J. (2001). Gaining and maintaining consent: Ethical concerns raised in a study of dying patients. *Qualitative Health Research*, *11*(5), 693–705. https://doi.org/10.1177/104973201129119389

Leh, A. (2000). Problems of archiving oral history interviews: The example of the archive "German Memory." *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *1*(3), Article 3. https://doi.org/10.17169/fqs-1.3.1025

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, *30*(4), 330–342. https://doi.org/10.1016/j.socnet.2008.07.002

Lewis, S. C. (2015). Journalism in an era of big data. *Digital Journalism*, *3*(3), 321–330. https://doi.org/10.1080/21670811.2014.976399

Liebowitz, J. (Ed.). (2013). *Big data and business analytics*. Auerbach Publications. https://doi.org/10.1201/b14700

Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M. E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D. V., Stockhause, M., & Westbrook, J. (2020). The TRUST Principles for digital repositories. *Scientific Data*, *7*(1), Article 144. https://doi.org/10.1038/s41597-020-0486-7

Lipset, S. M., & Bendix, R. (1959). *Social mobility in industrial society*. University of California Press.

Lipu, S., Williamson, K., & Lloyd, A. (2007). *Exploring methods in information literacy research*. Elsevier.

Lorentzen, D. G., & Nolin, J. (2017). Approaching completeness: Capturing a hashtagged Twitter conversation and its follow-on conversation. *Social Science Computer Review*, *35*(2), 277–286. https://doi.org/10.1177/0894439315607018

Lüders, M. (2008). Conceptualizing personal media. *New Media & Society*, *10*(5), 683–702. https://doi.org/10.1177/1461444808094352

Luo, L., & Wildemuth, B. M. (2017). Semistructured interviews. In *Applications of Social Research Methods to Questions in Information and Library Science* (2nd ed., pp. 248–257). Libraries Unlimited.

Madden, M. (2014). *Public perceptions of privacy and security in the post-Snowden era*. Pew Research Center. https://web.archive.org/web/20220415175921/https://www.pewresearch.org/internet/2014/11/12/public-privacy-perceptions/

Mannheimer, S. (2022). Data for: Connecting communities of practice: Data curation strategies for qualitative data reuse and big social research. *Qualitative Data Repository*. https://doi.org/10.5064/F6GWMU4O

Mannheimer, S., & Hull, E. A. (2018). Sharing selves: Developing an ethical framework for curating social media data. *International Journal of Digital Curation*, *12*(2), 196–209. https://doi.org/10.2218/ijdc.v12i2.518

Mannheimer, S., Pienta, A., Kirilova, D., Elman, C., & Wutich, A. (2019). Qualitative data sharing: Data repositories and academic libraries as key partners in addressing challenges. *American Behavioral Scientist*, *63*(5), 643–664. https://doi.org/10.1177/0002764218784991

Manovich, L. (2012). Trending: The promises and the challenges of big social data. In M. K. Gold (Ed.), *Debates in the Digital Humanities* (pp. 460–475). University of Minnesota Press. https://doi.org/10.5749/minnesota/9780816677948.003.0047

Markham, A. (2012). Fabrication as ethical practice. *Information, Communication & Society*, *15*(3), 334–353. https://doi.org/10.1080/1369118X.2011.641993

Martí, P., Serrano-Estrada, L., & Nolasco-Cirugeda, A. (2019). Social media data: Challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems*, *74*, 161–174. https://doi.org/10.1016/j.compenvurbsys.2018.11.001

Marwick, A. E., & boyd, d. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133. https://doi.org/10.1177/1461444810365313

Marwick, A. E., & boyd, d. (2014). Networked privacy: How teenagers negotiate context in social media. *New Media & Society*, *16*(7), 1051–1067. https://doi.org/10.1177/1461444814543995

Master, Z., & Resnik, D. B. (2013). Incorporating exclusion clauses into informed consent for biobanking. *Cambridge Quarterly of Healthcare Ethics*, *22*(2), 203–212. https://doi.org/10.1017/S0963180112000576

Mathur, A., Bleckman, J. D., & Lyle, J. (2017). Reuse of restricted-use research data. In

Curating research data, volume two: A handbook of current practice (pp. 258–261). Association of College and Research Libraries. http://deepblue.lib.umich.edu/handle/2027.42/135734

Mauthner, N. S. (2012). 'Accounting for our part of the entangled webs we weave': Ethical and moral issues in digital data sharing. In T. Miller, M. Birch, M. Mauthner, & J. Jessop, *Ethics in qualitative research* (pp. 157–175). SAGE Publications. https://doi.org/10.4135/9781473913912.n11

Mauthner, N. S., & Parry, O. (2009). Qualitative data preservation and sharing in the social sciences: On whose philosophical terms? *Australian Journal of Social Issues*, *44*(3), 291–307. https://doi.org/10.1002/j.1839-4655.2009.tb00147.x

Mauthner, N. S., & Parry, O. (2013). Open access digital data sharing: Principles, policies and practices. *Social Epistemology*, *27*(1), 47–67. https://doi.org/10.1080/02691728.2012.760663

Mauthner, N. S., Parry, O., & Backett-Milburn, K. (1998). The data are out there, or are they? Implications for archiving and revisiting qualitative data. *Sociology*, *32*(4), 733–745. https://doi.org/10.1177/0038038598032004006

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

McRory, W. (2021). Let the bots be bots: Why the CFAA must be clarified to prevent the selective banning of data collection facilitating private social media information monopolization. *Brooklyn Journal of Corporate, Financial & Commercial Law*, *16*(1), 279. https://brooklynworks.brooklaw.edu/bjcfcl/vol16/iss1/14

Mello, M. M., & Wolf, L. E. (2010). The Havasupai Indian Tribe case—Lessons for research involving stored biologic samples. *New England Journal of Medicine*, *363*(3), 204–207. https://doi.org/10.1056/NEJMp1005203

Metcalf, J. (2016). Big data analytics and revision of the common rule. *Communications of the ACM*, *59*(7), 31–33. https://doi.org/10.1145/2935882

Metcalf, J., & Crawford, K. (2016). Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society*, *3*(1), 1-14. https://doi.org/10.1177/2053951716650211

Michel, A., & Tappenbeck, I. (2019, May). *Information literacy, epistemic cultures and the question "Who needs what?"* Learning Information Literacy across the Globe, Frankfurt. https://doi.org/10.25656/01:17883

Miles, M. B., Huberman, A. M., & Saldana, J. (2020). *Qualitative data analysis: A methods*

sourcebook (4th ed.). SAGE Publications.

Mneimneh, Z., Pasek, J., Singh, L., Best, R., Bode, L., Bruch, E., Budak, C., Davis-Kean, P., Donato, K., Ellison, N., gelman, a., Groshen, E., Hemphill, L., Hobbs, W., Jensen, J. B., Karypis, G., Ladd, J. M., O'Hara, A., Raghunathan, T., … Wojcik, S. (2021). *Data Acquisition, Sampling, and Data Preparation Considerations for Quantitative Social Science Research Using Social Media Data*. https://doi.org/10.31234/osf.io/k6vyj

Moore, N. (2007). (Re)using qualitative data? *Sociological Research Online*, *12*(3), 1–13. https://doi.org/10.5153/sro.1496

Moore, N. (2012). The politics and ethics of naming: Questioning anonymisation in (archival) research. *International Journal of Social Research Methodology*, *15*(4), 331–340. https://doi.org/10.1080/13645579.2012.688330

Moreno, J. L. (1934). *Who shall survive?: A new approach to the problem of human interrelations.* Nervous and Mental Disease Publishing Co. https://doi.org/10.1037/10648-000

Mulder, A. E., Wiersma, G., & Loenen, B. V. (2020). Status of national open spatial data infrastructures: A comparison across continents. *International Journal of Spatial Data Infrastructures Research*, *15*, 56–87.

Murdock, G. P. (1961). *Outline of cultural materials*. Human Relations Area Files.

Mukurtu. (2020). *Mukurtu CMS*. https://web.archive.org/web/20220122150443/https://mukurtu.org/

Murray, J. B., & Evers, D. J. (1989). Theory borrowing and reflectivity in interdisciplinary fields. In T. K. Srull (Ed.), *NA - Advances in Consumer Research* (Vol. 16, pp. 647–652). The Association for Consumer Research. https://web.archive.org/web/20220406193140/https://www.acrwebsite.org/volumes/6929/volumes/v16/NA-16/full

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). The Belmont report. *United States Department of Health, Education, and Welfare*.

National Endowment for the Humanities. (2019). *Data management plans for NEH Office of Digital Humanities proposals and awards*. https://web.archive.org/web/20190906090545/https://www.neh.gov/sites/default/files/inline-files/Data%20Management%20Plans%2C%202019.pdf

National Institutes of Health. (2020). *Final NIH policy for data management and sharing*. https://web.archive.org/web/20220325154034/https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

National Science Foundation. (2011). *Dissemination and sharing of research results*. https://web.archive.org/web/20220327214059/http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Nazarenko, M. A., & Khronusova, T. V. (2017). Big data in modern higher education: Benefits and criticism. *Quality Management,Transport and Information Security, Information Technologies*, 676–679. https://doi.org/10.1109/ITMQIS.2017.8085914

Neale, B., & Bishop, L. (2012). The Timescapes Archive: A stakeholder approach to archiving qualitative longitudinal data. *Qualitative Research*, *12*(1), 53–65. https://doi.org/10.1177/1468794111426233

Nebeker, C., Dunseath, S. E., & Linares-Orozco, R. (2020). A retrospective analysis of NIH-funded digital health research using social media platforms: *Digital Health, 6,* 1-12. https://doi.org/10.1177/2055207619901085

Nippert-Eng, C. E. (2010). *Islands of privacy*. The University of Chicago Press.

Nissenbaum, H. (2009). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.

Obar, J. A., & Oeldorf-Hirsch, A. (2020). The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, *23*(1), 128–147. https://doi.org/10.1080/1369118X.2018.1486870

Oboler, A., Welsh, K., & Cruz, L. (2012). The danger of big data: Social media as computational social science. *First Monday*, *17*(7). https://doi.org/10.5210/fm.v17i7.3993

Office for Human Research Protections. (2018, July 30). *Revised Common Rule Q&As*. HHS.Gov. https://web.archive.org/web/20220302224303/https://www.hhs.gov/ohrp/education-and-outreach/revised-common-rule/revised-common-rule-q-and-a/index.html

Office of The Director. (2021). *NOT-OD-21-131: Request for information: Developing consent language for future use of data and biospecimens*. National Institutes of Health. https://web.archive.org/web/20211222113451/https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-131.html

Olshannikova, E., Olsson, T., Huhtamäki, J., & Kärkkäinen, H. (2017). Conceptualizing big social data. *Journal of Big Data*, *4*(1), Article 1. https://doi.org/10.1186/s40537-017-0063-x

Osterberg, G. (2017, December 26). Update on the Twitter archive at the Library of Congress [Blog]. *Library of Congress Blog*.

https://web.archive.org/web/20220405174129/https://blogs.loc.gov/loc/2017/12/update-on-the-twitter-archive-at-the-library-of-congress-2/

Otter.ai. (2021). *Otter.ai speech-to-text software*. https://web.archive.org/web/20220101184238/https://otter.ai/

Palen, L., & Dourish, P. (2003). Unpacking "privacy" for a networked world. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 129–136. https://doi.org/10.1145/642611.642635

Parry, O., & Mauthner, N. S. (2004). Whose data are they anyway? Practical, legal and ethical issues in archiving qualitative research data. *Sociology*, *38*(1), 139–152. https://doi.org/10.1177/0038038504039366

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and reuses of scientific data: The data creators' advantage. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.fc14bf2d

Paulus, T., Warren, A., & Lester, J. N. (2016). Applying conversation analysis methods to online talk: A literature review. *Discourse, Context & Media*, *12*, 1–10. https://doi.org/10.1016/j.dcm.2016.04.001

Petronio, S. S. (2002). *Boundaries of privacy: Dialectics of disclosure*. State University of New York Press.

Pettigrew, K. E., & McKechnie, L. (E F.). (2001). The use of theory in information science research. *Journal of the American Society for Information Science and Technology*, *52*(1), 62–73. https://doi.org/10.1002/1532-2890(2000)52:1<62::AID-ASI1061>3.0.CO;2-J

Pezalla, A. E., Pettigrew, J., & Miller-Day, M. (2012). Researching the researcher-as-instrument: An exercise in interviewer self-reflexivity. *Qualitative Research : QR*, *12*(2), 165–185. https://doi.org/10.1177/1487941111422107

Picciano, A. G. (2014). Big data and learning analytics in blended learning environments: Benefits and concerns. *IJIMAI*, *2*(7), 35–43. https://doi.org/10.9781/ijimai.2014.275

PLOS. (2014). *Data availability*. https://web.archive.org/web/20220502055901/https://journals.plos.org/plosone/s/data-availability

Proferes, N. (2017). Reaction to Cornelius Puschmann. In K. Kinder-Kurlanda & M. Zimmer (Eds.), *Internet Research Ethics for the Social Age* (p. 114). Peter Lang.

Pryse, J. A., Harp, M., Mannheimer, S., Marsolek, W., & Cowles, W. (2021). Oral history interviews data curation primer. *Data Curation Network*.

https://hdl.handle.net/11299/219052

Puebla, I., & Lowenberg, D. (2021). *Joint FORCE11 & COPE Research Data Publishing Ethics working group recommendations*. Zenodo. https://doi.org/10.5281/ZENODO.5391293

Puschmann, C. (2017). Bad judgment, bad ethics? Validity in computational social media research. In M. Zimmer & K. Kinder-Kurlanda (Eds.), *Internet Research Ethics for the Social Age* (pp. 95–113). Peter Lang.

Puschmann, C. (2019). An end to the Wild West of social media research: A response to Axel Bruns. *Information, Communication & Society*, *22*(11), 1582–1589. https://doi.org/10.1080/1369118X.2019.1646300

Puschmann, C., & Burgess, J. (2014). Metaphors of big data. *International Journal of Communication*, *8*. https://ijoc.org/index.php/ijoc/article/view/2169

QSR International. (2022). *NVivo qualitative data analysis software*. https://web.archive.org/web/20220402173708/https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems*, *2*(1), Article 3. https://doi.org/10.1186/2047-2501-2-3

Raguseo, E. (2018). Big data technologies: An empirical investigation on their adoption, benefits and risks for companies. *International Journal of Information Management*, *38*(1), 187–195. https://doi.org/10.1016/j.ijinfomgt.2017.07.008

Rains, S. A., & Brunner, S. R. (2015). What can we learn about social network sites by studying Facebook? A call and recommendations for research on social network sites. *New Media & Society*, *17*(1), 114–131. https://doi.org/10.1177/1461444814546481

Ramasamy, D., Venkateswaran, S., & Madhow, U. (2013). Inferring user interests from tweet times. *Proceedings of the First ACM Conference on Online Social Networks*, 235–240. https://doi.org/10.1145/2512938.2512960

Re3data. (2019). *Registry of Research Data Repositories*. https://doi.org/10.17616/R3D

Resnick, B. (2016, May 12). Researchers just released profile data on 70,000 OkCupid users without permission. *Vox*. https://web.archive.org/web/20220221130935/https://www.vox.com/2016/5/12/11666116/70000-okcupid-users-data-release

Reuter, K., Zhu, Y., Angyan, P., Le, N., Merchant, A. A., & Zimmer, M. (2019). Public concern about monitoring Twitter users and their conversations to recruit for clinical trials:

Survey study. *Journal of Medical Internet Research*, *21*(10), e15455. https://doi.org/10.2196/15455

Rivers, C. M., & Lewis, B. L. (2014). Ethical research standards in a world of big data. *F1000Research*. https://doi.org/10.12688/f1000research.3-38.v2

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed). Free Press.

Rothstein, M. A. (2010). Is deidentification sufficient to protect health privacy in research? *The American Journal of Bioethics*, *10*(9), 3–11. https://doi.org/10.1080/15265161.2010.494215

Ruthven, I., Buchanan, S., & Jardine, C. (2018). Relationships, environment, health and development: The information needs expressed online by young first-time mothers. *Journal of the Association for Information Science and Technology*, *69*(8), 985–995. https://doi.org/10.1002/asi.24024

Schaeffer, P. (2011, October 5). *Why does Dryad use CC0?* Dryad News and Views. https://web.archive.org/web/20220401073936/https://blog.datadryad.org/2011/10/05/why-does-dryad-use-cc0/

Scheff, T. J. (1986). Toward resolving the controversy over "thick description." *Current Anthropology*, *27*(4), 408–409. https://doi.org/10.1086/203460

Schema.org. (2020). *Data and datasets*. https://web.archive.org/web/20211215014211/https://schema.org/docs/data-and-datasets.html

Schneble, C. O., Elger, B. S., & Shaw, D. (2018). The Cambridge Analytica affair and internet-mediated research. *EMBO Reports*, *19*(8), e46579. https://doi.org/10.15252/embr.201846579

Schöch, C. (2017). Wiederholende forschung in den digitalen geisteswissenschaften. *Digital Nachhaltigkeit (DHd2017),* 207-212. https://doi.org/10.5281/zenodo.277113

Schofield, P. N., Bubela, T., Weaver, T., Portilla, L., Brown, S. D., Hancock, J. M., Einhorn, D., Tocchini-Valentini, G., Hrabe de Angelis, M., & Rosenthal, N. (2009). Post-publication sharing of data and tools. *Nature*, *461*(7261), 171–173. https://doi.org/10.1038/461171a

Schreier, M. (2018). Sampling and generalization. In U. Flick (Ed.), *The SAGE handbook of qualitative data collection* (pp. 84–97). SAGE Publications. https://doi.org/10.4135/9781526416070.n6

Schroeder, R. (2016). Big data business models: Challenges and opportunities. *Cogent Social Sciences*, *2*(1). https://doi.org/10.1080/23311886.2016.1166924

Secretary's Advisory Committee on Human Research Protections. (2013). *Attachment B: Considerations and recommendations concerning internet research and human subjects research regulations, with revisions*. https://web.archive.org/web/20220119141810/https://www.hhs.gov/ohrp/sites/default/files/ohrp/sachrp/mtgings/2013%20March%20Mtg/internet_research.pdf

Secretary's Advisory Committee on Human Research Protections. (2015). *Attachment A: Human subjects research implications of "big data."* https://web.archive.org/web/20220409135614/https://www.hhs.gov/ohrp/sachrp-committee/recommendations/2015-april-24-attachment-a/index.html

Segerberg, A., & Bennett, W. L. (2011). Social media and the organization of collective action: Using Twitter to explore the ecologies of two climate change protests. *The Communication Review*, *14*(3), 197–215. https://doi.org/10.1080/10714421.2011.597250

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science: Possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 6–13. https://doi.org/10.1177/0002716215572084

Sherif, V. (2018). Evaluating preexisting qualitative research data for secondary analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *19*(2), Article 2. https://doi.org/10.17169/fqs-19.2.2821

Shilton, K., & Sayles, S. (2016). "We aren't all going to be on the same page about ethics": Ethical practices and challenges in research on digital and social media. *49th Hawaii International Conference on System Sciences (HICSS)*, 1909–1918. https://doi.org/10.1109/HICSS.2016.242

Sieber, J. E. (1991a). Introduction: Sharing social science data. In J. E. Sieber (Ed.), *Sharing social science data: Advantages and challenges* (pp. 1–18). SAGE Publications.

Sieber, J. E. (1991b). Social scientists' concerns about sharing data. In J. E. Sieber (Ed.), *Sharing social science data: Advantages and challenges* (pp. 141–150). SAGE Publications.

Siminoff, L. A. (2003). Toward improving the informed consent process in research with humans. *IRB: Ethics & Human Research*, *25*(5), S1–S3. https://doi.org/10.2307/3564115

Simmel, G. (1955). *Conflict and the web of group affiliations*. The Free Press.

Simons, N., Goodey, G., Hardeman, M., Clare, C., Gonzales, S., Strange, D., Smith, G., Kipnis, D., Iida, K., Miyairi, N., Tshetsha, V., Ramokgola, R., Makhera, P., & Barbour, G. (2021).

The state of open data 2021. *Digital Science Report*. https://doi.org/10.6084/m9.figshare.17061347.v1

Sleeper, M., Balebako, R., Das, S., McConahy, A. L., Wiese, J., & Cranor, L. F. (2013). The post that wasn't: Exploring self-censorship on Facebook. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*, 793–802. https://doi.org/10.1145/2441776.2441865

Sloan, L. (2016). Social science 'lite'? Deriving demographic proxies from Twitter. In L. Sloan & A. Quan-Haase, *The SAGE handbook of social media research methods* (pp. 90–104). SAGE Publications. https://doi.org/10.4135/9781473983847.n7

Smith, E. (2008). *Using Secondary Data in Educational and Social Research*. McGraw-Hill Education.

Smith, P. L., Felima, C., Durant, F., Van Kleeck, D., Huet, H., & Taylor, L. N. (2020). Building socio-technical systems to support data management and digital scholarship in the social sciences. In J. W. Crowder, M. Fortun, R. Besara, & L. Poirier (Eds.), *Anthropological data in the digital age: New possibilities—New challenges* (pp. 31–57). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-24925-0_3

State of California. (2020). *Title 11. Law—Division 1. Attorney General—Chapter 20. California Consumer Privacy Act Regulations*. https://web.archive.org/web/20220401105842/https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5

State of Vermont. (2018). *H.764 (Act 171)*. https://web.archive.org/web/20220403105500/https://legislature.vermont.gov/Documents/2018/Docs/ACTS/ACT171/ACT171%20As%20Enacted.pdf

Stenbacka, C. (2001). Qualitative research requires quality concepts of its own. *Management Decision*, *39*(7), 551–556. https://doi.org/10.1108/EUM0000000005801

Stevens, M., Wehrens, R., & de Bont, A. (2020). Epistemic virtues and data-driven dreams: On sameness and difference in the epistemic cultures of data science and psychiatry. *Social Science & Medicine*, *258*, 113116. https://doi.org/10.1016/j.socscimed.2020.113116

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, *38*(5), 503–516. https://doi.org/10.1177/0894439319843669

Stoycheff, E., Liu, J., Wibowo, K. A., & Nanni, D. P. (2017). What have we learned about social media by studying Facebook? A decade in review. *New Media & Society*, *19*(6), 968–980. https://doi.org/10.1177/1461444817695745

Sujon, Z. (2017). Reaction to Tromble and Stockmann. In K. Kinder-Kurlanda & M. Zimmer (Eds.), *Internet Research Ethics for the Social Age*. Peter Lang.

Summers, E. (2017, August 21). *The catalog and the hydrator*. Documenting the Now. https://news.docnow.io/the-catalog-and-the-hydrator-3299eddfe21e

Sweeney, L., Crosas, M., & Bar-Sinai, M. (2015). Sharing sensitive data with confidence: The datatags system. *Technology Science*, 2015101601. https://web.archive.org/web/20220122022200/https://techscience.org/a/2015101601/

Szabo, V., & Strang, V. R. (1997). Secondary analysis of qualitative data. *Advances in Nursing Science*, *20*(2), 66. https://doi.org/10.1097/00012272-199712000-00008

Taichman, D. B., Sahni, P., Pinborg, A., Peiperl, L., Laine, C., James, A., Hong, S.-T., Haileamlak, A., Gollogly, L., Godlee, F., Frizelle, F. A., Florenzano, F., Drazen, J. M., Bauchner, H., Baethge, C., & Backus, J. (2017). Data sharing statements for clinical trials. *BMJ*, *357,* j2372. https://doi.org/10.1136/bmj.j2372

Talja, S., Tuominen, K., & Savolainen, R. (2005). "Isms" in information science: Constructivism, collectivism and constructionism. *Journal of Documentation*, *61*(1), 79–101. https://doi.org/10.1108/00220410510578023

TEI Consortium. (2019). *TEI: Text Encoding Initiative*. https://web.archive.org/web/20220401002343/https://tei-c.org/

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLOS ONE*, *6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Tenopir, C., Sandusky, R. J., Allard, S., & Birch, B. (2014). Research data management services in academic research libraries and perceptions of librarians. *Library & Information Science Research*, *36*(2), 84–90. https://doi.org/10.1016/j.lisr.2013.11.003

Tenopir, C., Talja, S., Horstmann, W., Late, E., Hughes, D., Pollock, D., Schmidt, B., Baird, L., Sandusky, R. J., & Allard, S. (2017). Research data services in European academic research libraries. *LIBER Quarterly*, *27*(1), 23–44. https://doi.org/10.18352/lq.10180

The Economist. (2022, January 29). Your secret's safe with me; Data privacy. *The Economist*, 62–63.

Thompson, P. (2000). Re-using qualitative research data: A personal account. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *1*(3), Article 3. https://doi.org/10.17169/fqs-1.3.1044

Thomson, S. D. (2016). Preserving social media. *Digital Preservation Coalition Technology*

Watch Report. https://doi.org/10.7207/twr16-01

Thorne, S. (1994). Secondary analysis in qualitative research: Issues and implications. In J. M. Morse (Ed.), *Critical issues in qualitative research methods* (pp. 263–279). SAGE Publications.

Thorne, S. (1998). Ethical and representational issues in qualitative secondary analysis. *Qualitative Health Research*, *8*(4), 547–555. https://doi.org/10.1177/104973239800800408

Thorne, S. (2004). Secondary analysis of qualitative data. In M. Lewis-Beck, A. Bryman, & T.F. Liao (Eds.), *The SAGE encyclopedia of social science research methods*. SAGE Publications. https://dx.doi.org/10.4135/9781412950589.n895

Tiffin, N. (2018). Tiered informed consent: Respecting autonomy, agency and individuality in Africa. *BMJ Global Health*, *3*(6), e001249. https://doi.org/10.1136/bmjgh-2018-001249

Title 17. Copyrights. Chapter 1. Subject matter and scope of copyright., Cornell Law School Legal Information Institute (1990). https://web.archive.org/web/20220401152536/https://www.law.cornell.edu/uscode/text/17/102

Törnberg, P., & Törnberg, A. (2018). The limits of computation: A philosophical critique of contemporary Big Data research. *Big Data & Society*, *5*(2), 1-12. https://doi.org/10.1177/2053951718811843

Tsai, A. C., Kohrt, B. A., Matthews, L. T., Betancourt, T. S., Lee, J. K., Papachristos, A. V., Weiser, S. D., & Dworkin, S. L. (2016). Promises and pitfalls of data sharing in qualitative research. *Social Science & Medicine*, *169*, 191–198. https://doi.org/10.1016/j.socscimed.2016.08.004

Turnbull, A. (2000). Collaboration and censorship in the oral history interview. *International Journal of Social Research Methodology*, *3*(1), 15–34. https://doi.org/10.1080/136455700294905

Twitter. (2020). *Developer terms. Developer agreement and policy*. https://web.archive.org/web/20220402015546/https://developer.twitter.com/en/developer-terms/agreement-and-policy

Twitter. (2022, March 2). *Honoring user intent on Twitter*. Twitter Developer Platform. https://web.archive.org/web/20220302220014/https://developer.twitter.com/en/docs/twitter-api/enterprise/compliance-firehose-api/guides/honoring-user-intent

Twitter Developers. (2020). *More about restricted uses of the Twitter APIs*. https://web.archive.org/web/20220426002748/https://developer.twitter.com/en/de

veloper-terms/more-on-restricted-use-cases

UK Data Archive. (2019). *Metadata standards: QuDEx*. https://web.archive.org/web/20220302004944/https://www.data-archive.ac.uk/managing-data/standards-and-procedures/metadata-standards/

UK Data Archive. (2022). *QuDEx*. UK Data Archive - University of Essex. https://web.archive.org/web/20220119095028/https://www.data-archive.ac.uk/managing-data/standards-and-procedures/metadata-standards/qudex/

U.S. Department of Health and Human Services. (1991). Federal policy for the protection of human subjects ("Common rule"). In *HHS.gov*. https://web.archive.org/web/20220401062540/https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html

Van Buren v. United States, 141 Supreme Court Reporter 1648 (Supreme Court 2021). https://web.archive.org/web/20220324061739/https://www.supremecourt.gov/opinions/20pdf/19-783_k53l.pdf

van de Sandt, S., Dallmeier-Tiessen, S., Lavasa, A., & Petras, V. (2019). The definition of reuse. *Data Science Journal*, *18*, 22. https://doi.org/10.5334/dsj-2019-022

van den Berg, H. (2005). Reanalyzing qualitative interviews from different angles: The risk of decontextualization and other problems of sharing qualitative data. *Historical Social Research / Historische Sozialforschung*, *6*(1), Article 30. https://doi.org/10.17169/fqs-6.1.499

VandeVusse, A., Mueller, J., & Karcher, S. (2022). Qualitative data sharing: Participant understanding, motivation, and consent. *Qualitative Health Research*, *32*(1), 182–191. https://doi.org/10.1177/10497323211054058

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*, 280-289. https://web.archive.org/web/20220307114705/https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587/14817

Verma, I. M. (2014). Editorial expression of concern: Experimental evidence of massivescale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(29), 10779–10779. https://doi.org/10.1073/pnas.1412469111

Vestoso, M. (2018). The GDPR beyond privacy: Data-driven challenges for social scientists, legislators and policy-makers. *Future Internet*, *10*(7), 62. https://doi.org/10.3390/fi10070062

Viceconti, M., Hunter, P., & Hose, R. (2015). Big data, big knowledge: Big data for

personalized healthcare. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1209–1215. https://doi.org/10.1109/JBHI.2015.2406883

Villarroel Ordenes, F., Grewal, D., Ludwig, S., Ruyter, K. D., Mahr, D., & Wetzels, M. (2019). Cutting through content clutter: How speech and image acts drive consumer sharing of social media brand messages. *Journal of Consumer Research*, *45*(5), 988–1012. https://doi.org/10.1093/jcr/ucy032

Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., & Mechant, P. (2019). Web archives as a data resource for digital scholars. *International Journal of Digital Humanities*, *1*(1), 85–111. https://doi.org/10.1007/s42803-019-00007-7

Voigt, P., & von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR)*. Springer International Publishing. https://doi.org/10.1007/978-3-319-57959-7

Walsh, C. (2020, August 11). Challenge of archiving the #MeToo movement. *Harvard Gazette*. https://web.archive.org/web/20220105190819/https://news.harvard.edu/gazette/story/2020/08/challenge-of-archiving-the-metoo-movement/

Walters, P. (2009). Qualitative archiving: Engaging with epistemological misgivings. *Australian Journal of Social Issues*, *44*(3), 309–320. https://doi.org/10.1002/j.1839-4655.2009.tb00148.x

Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, *126*, 3–13. https://doi.org/10.1016/j.techfore.2015.12.019

Washington Post. (2018, April 10). *Transcript of Mark Zuckerberg's Senate hearing*. https://web.archive.org/web/20220410013853/https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-zuckerbergs-senate-hearing/

Wellcome. (2017). *Data, software and materials management and sharing policy*. https://web.archive.org/web/20220121022013/https://wellcome.org/grant-funding/guidance/data-software-materials-management-and-sharing-policy

Weller, K., & Kinder-Kurlanda, K. E. (2016). A manifesto for data sharing in social media research. *Proceedings of the 8th ACM Conference on Web Science (WebSci '16)*, 166–172. https://doi.org/10.1145/2908131.2908172

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity.* Cambridge University Press.

Wenger, E., McDermott, R. A., & Snyder, W. (2002). *Cultivating communities of practice: A guide to managing knowledge*. Harvard Business School Press.

White, D. R. (1991). Sharing anthropological data with peers and Third World hosts. In J. E. Sieber (Ed.), *Sharing social science data: Advantages and challenges* (pp. 42–60). SAGE Publications.

Wildemuth, B. M. (Ed.). (2017). *Applications of social research methods to questions in information and library science* (2nd ed.). Libraries Unlimited.

Wiles, R., Crow, G., Charles, V., & Heath, S. (2007). Informed consent and the research process: Following rules or striking balances? *Sociological Research Online*, *12*(2), 1–12. https://doi.org/10.5153/sro.1208

Wilkinson, D., & Thelwall, M. (2011). Researching personal information on the public web: Methods and ethics. *Social Science Computer Review*, *29*(4), 387–401. https://doi.org/10.1177/0894439310378979

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Wilkof, N. (2016). IP knowledge in the age of Wikipedia and the blogosphere. *Journal of Intellectual Property Law & Practice*, *11*(7), 477–478. https://doi.org/10.1093/jiplp/jpw072

Williams, S. A., Terras, M. M., & Warwick, C. (2013). What do people study when they study Twitter? Classifying Twitter related academic papers. *Journal of Documentation*, *69*(3), 384–410. https://doi.org/10.1108/JD-03-2012-0027

Williamson, B. (2017). *Big data in education: The digital future of learning, policy and practice.* SAGE Publications.

Wilson, D. W., Lin, X., Longstreet, P., & Sarker, S. (2011). *Web 2.0: A definition, literature review, and directions for future research*.

Wilson, R. E., Gosling, S. D., & Graham, L. T. (2012). A review of Facebook research in the social sciences. *Perspectives on Psychological Science*, *7*(3), 203–220. https://doi.org/10.1177/1745691612442904

Winskell, K., Singleton, R., & Sabben, G. (2018). Enabling analysis of big, thick, long, and wide data: Data management for the analysis of a large longitudinal and cross-national narrative data set. *Qualitative Health Research*. https://doi.org/10.1177/1049732318759658

Wittwer, M., Reinhold, O., & Alt, R. (2017). Capturing customer context from social media: Mapping social media API and CRM profile data. *Proceedings of the International*

Conference on Web Intelligence, 993–997. https://doi.org/10.1145/3106426.3117762

World Medical Association. (2013). *Declaration of Helsinki*. https://web.archive.org/web/20220430103920/https://www.wma.net/what-we-do/medical-ethics/declaration-of-helsinki/

Wutich, A., & Brewis, A. (2014). Food, water, and scarcity: Toward a broader anthropology of resource insecurity. *Current Anthropology*, *55*(4), 444–468. https://doi.org/10.1086/677311

Yanai, K. (2012). World seer: A realtime geo-tweet photo mapping system. *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, 1–2. https://doi.org/10.1145/2324796.2324870

Yoon, A. (2017). Data reusers' trust development. *Journal of the Association for Information Science and Technology*, *68*(4), 946–956. https://doi.org/10.1002/asi.23730

Yoon, A., & Lee, Y. Y. (2019). Factors of trust in data reuse. *Online Information Review*. https://doi.org/10.1108/OIR-01-2019-0014

Yoon, A., & Schultz, T. (2017). Research data management services in academic libraries in the U.S.: A content analysis of libraries' websites. *College & Research Libraries*, *78*(7). https://doi.org/10.5860/crl.78.7.920

Zhang, W., Johnson, T. J., Seltzer, T., & Bichard, S. L. (2010). The revolution will be networked: The influence of social networking sites on political attitudes and behavior. *Social Science Computer Review*, *28*(1), 75–92. https://doi.org/10.1177/0894439309335162

Zhang, Y., & Wildemuth, B. M. (2017). Qualitative analysis of content. In *Applications of Social Research Methods to Questions in Information and Library Science* (2nd ed.). Libraries Unlimited.

Zhang, Z., He, Q., Gao, J., & Ni, M. (2018). A deep learning approach for detecting traffic accidents from social media data. *Transportation Research Part C: Emerging Technologies*, *86*, 580–596. https://doi.org/10.1016/j.trc.2017.11.027

Zikopoulos, P. (2012). *Understanding big data: Analytics for enterprise class Hadoop and streaming data*. McGraw-Hill.

Zimmer, M. (2010). "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology*, *12*(4), 313–325. https://doi.org/10.1007/s10676-010-9227-5

Zimmer, M. (2015). The Twitter Archive at the Library of Congress: Challenges for information practice and information policy. *First Monday*.

https://doi.org/10.5210/fm.v20i7.5619

Zimmer, M. (2016, May 14). OkCupid study reveals the perils of big-data science. *Wired*. https://web.archive.org/web/20220423184354/https://www.wired.com/2016/05/okcupid-study-reveals-perils-big-data-science/

Zimmer, M. (2018). Addressing conceptual gaps in big data research ethics: An application of contextual integrity. *Social Media + Society*, *4*(2), Article 2. https://doi.org/10.1177/2056305118768300

Zimmer, M., & Proferes, N. J. (2014). A topology of Twitter research: Disciplines, methods, and ethics. *Aslib Journal of Information Management*, *66*(3), 250–261. https://doi.org/10.1108/AJIM-09-2013-0083

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. https://doi.org/10.1177/0162243907306704

Zoom. (2021). *Video conferencing, cloud phone, webinars, chat, virtual events*. https://web.archive.org/web/20220102000820/https://zoom.us/

Appendix 1. Consent agreement for interviews, v1

Consent for participation in human research at Montana State University

Project Title

Connecting communities of practice: Using strategies from qualitative data curation to support big social research

Introduction

Request

You are being asked to participate in a research interview discussing your experience curating and/or conducting research with qualitative data and/or big social data.

Outcome

This study will inform the development of strategies to support ethical, legal, and epistemologically-sound qualitative and big social research.

Reasoning

I have identified potential participants by reviewing relevant conferences and journals. Additional participants have been identified through snowball sampling.

Procedure

Participation is voluntary and there is no cost to you to participate. If you agree to participate, you will be asked to discuss your experience curating or conducting research with qualitative data and/or big social data.

Risks and Benefits

There are no foreseen risks to participating in the study. The study is of no direct benefit to you.

Decline to Participate

You may decline to participate, and you may withdraw at any time.

Study Funding

There is no declared funding.

Confidentiality

Interviews will be recorded. All data will be stored securely during collection. Excerpts from your interview may be published, with your personal information deidentified. Full interview transcripts and qualitative analysis will be deidentified and published in a data repository.

Questions or Concerns

If you have any questions about this project, please contact Sara Mannheimer, 907-223-6323, sara.mannheimer@montana.edu. If you have additional questions about the rights of human subjects, you may contact the Chair of the Institutional Review Board, Mark Quinn,
406- 994-4707, mquinn@montana.edu.

AUTHORIZATION: I have read the above and understand the discomforts, inconveniences and risk of this study. I, _____, agree to participate in this research.

I understand that I may later refuse to participate, and that I may withdraw from the study at any time.

Signature of Participant: _____Date: _____

Signature of Investigator: _____Date: _____

Appendix 2. Consent agreement for interviews, v2

Consent for participation in human research at Montana State University

Project Title

Connecting communities of practice: Using strategies from qualitative data curation to support big social research

Introduction

Request: You are being asked to participate in a research interview discussing your experience curating and/or conducting research with qualitative data and/or big social data. **Outcome:** This study will inform the development of strategies to support ethical, legal, and epistemologically sound qualitative and big social research. **Sampling:** I have identified potential participants by reviewing relevant journal articles and datasets. Additional participants have been identified through snowball sampling.

Procedure

Participation is voluntary and there is no cost to you to participate. If you agree to participate, you will be asked to discuss your experience curating or conducting research with qualitative data and/or big social data.

Risks and Benefits

There are no major risks to participating in the study, but you will be discussing issues you encounter in your work and research practices. The study is of no direct benefit to you.

Decline to Participate

You may decline to participate, and you may withdraw at any time.

Study Funding

There is no declared funding.

Confidentiality

Interviews will be recorded. All data will be stored securely during collection. Excerpts from your interview may be published, with your personal information deidentified.

Questions or Concerns

If you have any questions about this project, please contact Sara Mannheimer, 907-223-6323, sara.mannheimer@montana.edu. If you have additional questions about the rights of human subjects, you may contact the Chair of the Institutional Review Board, Mark Quinn,
406- 994-4707, mquinn@montana.edu.

AUTHORIZATION: I have read the above and understand the discomforts, inconveniences and risks of this study. I, _____, agree to participate in this research. I understand that I may later refuse to participate, and that I may withdraw from the study at any time.

I agree to allow the deidentified transcript from my interview to be published in a data repository. ▢ YES ▢ NO

Signature of Participant: _____Date: _____

Signature of Investigator: _____Date: _____

Appendix 3. Qualitative researchers interview guide

Informed consent

We will review and sign the consent form.

Project Overview

Project Title

Connecting communities of practice: Using strategies from qualitative data curation to support big social research

Research summary

Big social data (such as social media and blogs) and archived qualitative data (such as interview transcripts, field notebooks, and diaries) are similar, but their respective communities of practice are under-connected. Research with both types of data repurpose existing social data to advance discoveries in social science. However, despite these similarities, big social research has not yet been widely framed as a form of qualitative data reuse, and qualitative data reuse has only begun to be discussed through a big social data lens. Qualitative data reuse is a more established practice, and therefore has more developed data curation strategies to support data sharing. My research investigates how data curation practices from each of these communities can inform the other for mutual benefit. The research will use interviews of qualitative researchers, big social data researchers, and data curators to gain insights into different community approaches to research and data sharing.

Research background

This research asks: how can data curators best handle qualitative and big social data to support ethical, epistemological, and legal data sharing practices?

My review of the literature revealed that there are six key issues that pose challenges for both groups. During the interview, I will ask questions about your personal experience in each of these six topic areas, plus introductory and wrap up questions. The interview will take 60-75 minutes.

Thank you for taking the time to speak with me! Your interview will help to improve data curation practices across disciplines.

Interview questions

We'll start recording the interview now.

Introductory question

Tell me about the type of research you do and what kind of data you generally produce.

Identifying a specific example

Describe a recent time when you:

- prepared your qualitative data for publication or sharing; or
- reused existing qualitative data yourself; or
- considered sharing your qualitative data, even if you ended up deciding against sharing; or
- observed firsthand someone else doing one of the above.

Was this example part of a grant-funded project that required specific treatment of the data?

Did you have a data management plan?

If you published any of the data from your example:

- Did you publish in a repository? Which one?
- What are your plans for storing, retaining, and deleting data in the future?
- Who has access to the data?

Context

> Qualitative research is a process that may include deep and prolonged contact and connection with research subjects, attempting to understand subjects within their own context (Miles et al., 2020). Qualitative data are therefore highly context-dependent. As Hinds et al. write, "context is a source of data, meaning, and understanding… Ignoring context, underusing it, or not recognizing one's own context-driven perspective will result in incomplete or missed meaning and a misunderstanding of human phenomena" (1992, p. 72).

1. Tell me about a time (if any) during your example when you considered the issue of understanding, maintaining, or communicating the data's context (e.g. contextual information about the community where the data was collected, contextual information about respondents)?

- What were your considerations? What concerns did you have? What factors helped you better understand/communicate context? What factors prevented you from understanding/communicating context?
- Was your research affected by incomplete contextual information?

- Did you consult with anyone, consider other research projects, or refer to literature, policies, or guidelines regarding this issue? Please explain.
- What strategies did you use to discern/communicate context during your example?
- If your example includes publishing your data, what strategies did you use to communicate context to potential future users of the data?

Data quality

2. During your example, what quality issues or concerns arose (for example, missing data, bias, or quality of method)?

- What factors helped you better understand/communicate data quality issues? What factors prevented you from understanding/communicating data quality?
- Did you consult with anyone, consider other research projects, or refer to literature, policies, or guidelines regarding this issue? Please explain.
- What strategies did you use (or did you see used) to communicate, describe, or clarify data quality issues in your example? Can you describe in detail how those strategies helped you?

Data comparability

3. During your example, did you compare and/or combine multiple qualitative datasets?

- Or: did you consider comparability or interoperability of your dataset?

If no:

- why not?

If yes:

- Why (for what purpose) did you combine the datasets? How did this advance your research?
- Did you consult with anyone, consider other research projects, or refer to literature, policies, or guidelines regarding this issue? Please explain.
- What challenges did you encounter when combining multiple qualitative datasets, and what strategies did you use to address these challenges?

Informed consent

4. Tell me about a time (if any) during the process of your example when you considered the idea of consent, particularly consent for future use of the data.

- For examples involving data reuse
    - Did you consider consent from original respondents when conducting your research?
- For all examples

- Did you consult with anyone, consider other research projects, or refer to literature, policies, or guidelines regarding this issue? Please explain.
- What strategies did you use to support consent for future use of your data (e.g., broad consent, focus groups, community advisory boards)?

Privacy and confidentiality

5. Tell me about a time (if any) during the process of your example when you considered issues of privacy and confidentiality.

- During your example, what do you think the participants' expectations of privacy were?
- Did you consult with anyone, consider other research projects, or refer to literature, policies, or guidelines regarding this issue? Please explain.
- What strategies did you use to address privacy (e.g. restricted access, de-identification)?
- Did you feel that you had to make any compromises about participant privacy in order to publish your data/conduct your secondary research? If so, how so?

Intellectual property

6. Tell me about a time (if any) during the process of your example when you considered intellectual property concerns (especially if you published your data or reused existing data)? (e.g. participant intellectual property, organizational IP, the idea of Fair Use)

- Did you consult with anyone, consider other research projects, or refer to literature, policies, or guidelines regarding this issue? Please explain.
- What strategies did you use to address the issue of intellectual property?

Additional issues

7. Are there issues or challenges that arose during your example that I haven’t asked you about?

8. Who else should I interview? I’m looking for big social researchers, qualitative researchers who have published or reused data, and data curators who have worked with qualitative or big social data.

Appendix 4. Big social researchers interview guide

Informed consent

We will review and sign the consent form.

Project Overview

Project Title

Connecting communities of practice: Using strategies from qualitative data curation to support big social research

Research summary

Big social data (such as social media and blogs) and archived qualitative data (such as interview transcripts, field notebooks, and diaries) are similar, but their respective communities of practice are under-connected. Research with both types of data repurpose existing social data to advance discoveries in social science. However, despite these similarities, big social research has not yet been widely framed as a form of qualitative data reuse, and qualitative data reuse has only begun to be discussed through a big social data lens. Qualitative data reuse is a more established practice, and therefore has more developed data curation strategies to support data sharing. My research investigates how data curation practices from each of these communities can inform the other for mutual benefit. The research will use interviews of qualitative researchers, big social data researchers, and data curators to gain insights into different community approaches to research and data sharing.

Research background

This research asks: how can data curators best handle qualitative and big social data to support ethical, epistemological, and legal data sharing practices?

My review of the literature revealed that there are six key issues that pose challenges for both groups. During the interview, I will ask questions about your personal experience in each of these six topic areas, plus introductory and wrap up questions. The interview will take 60-75 minutes.

Thank you for taking the time to speak with me! Your interview will help to improve data curation practices across disciplines.

Interview questions

We'll start recording the interview now.

Introductory question

Tell me about the type of research you do and what kind of data you produce.

Identifying a specific example

Please describe a recent time when you:

- collected big social data for research; or
- reused big social data that was collected and shared by someone else; or
- prepared big social data for publication or sharing.

Please also describe your data collection method (API, scraping, shared dataset, etc.)

Was this example part of a grant-funded project that required specific treatment of the data?

Did you have a data management plan?

Did you publish any of the data from your example?

- Is the data published in a repository? Which one?
- What are your plans for storing, retaining, and deleting data in the future?
- Who has access to the data?

Context

> Halavais (2015) suggests that "when we collect data from [social media] platforms (just as when we collected data in traditional spaces), context matters." However, the context of a social media post may be absent or difficult to understand. Social media posts are by nature short pieces of text, images, videos, etc, taken from a larger context of personal and public life. This out-of-context effect is only compounded when data are amassed at a large scale.

1. Tell me about a time (if any) during the process of your example when you considered the issue of maintaining and understanding the data's context (i.e. contextual information about the community where the data was collected, contextual information about respondents)?

- What were your considerations? What concerns did you have? What factors helped you better understand/communicate context? What factors prevented you from understanding/communicating context? Was your research affected by incomplete contextual information?
- Did you consult with anyone, consider other research projects, or refer to literature, policies, or guidelines regarding this issue? Please explain.

- What strategies did you use to discern/understand context during your example?
- If the example includes publishing your own big social data, what strategies did you use to communicate context to future users?

Data quality

2. Tell me about a time (if any) during the process of your example when you considered the issue of data quality (for example, missing data, bots, bias, quality of method)?

- How did you assess quality? What data quality issues arose?
- What data quality concerns did you have?
- What factors helped you better consider data quality issues? What factors prevented you from considering data quality?
- Did you consult with anyone, consider other research projects, or refer to literature, policies, or guidelines regarding this issue? Please explain.
- What strategies did you use (or did you see used) to communicate, describe, or clarify data quality issues in your example? Please describe in detail how those strategies helped you.

Data comparability

3. During your example, did you compare and/or combine multiple big social datasets?

- or: Did you consider comparability or interoperability of your dataset?
- Did you consult with anyone, consider other research projects, or refer to literature, policies, or guidelines regarding this issue? Please explain.

If no:

- why not?

If yes:

- Why (for what purpose) did you combine the datasets? How did this advance your research?
    - what dataset did you combine it with - where did that data come from? Your own or someone else's?
- What strategies did you use to combine multiple qualitative datasets?
    - what challenges did you encounter and how did you address them?

Informed consent

4. Tell me about a time (if any) during the process of your example when you considered informed consent.

- Did you consult with anyone, consider other research projects, or refer to literature, policies, or guidelines regarding this issue (including IRB)? Please explain.

- Have you used any other type of consent besides informed consent per se (e.g. broad consent, focus groups, community advisory boards)
- Did you feel that participants in your research would expect to give informed consent for the research?

Privacy and confidentiality

5. Tell me about a time (if any) during the process of your example when you considered issues of privacy (e.g. protecting data during research, considering restricted access or TweetIDs only if publishing).

- Did you consult with anyone, consider other research projects, or refer to literature, policies, or guidelines regarding this issue? Please explain.
- During your example, what do you think the participants' expectations of privacy were?
- Did you feel that you had to make any compromises about participant privacy in order to conduct your research?
- What strategies did you use to address the issue of privacy?

Intellectual property

6. Tell me about a time (if any) during the process of your example when you considered intellectual property concerns when you conducted your research and/or published your data (e.g. social media platform terms of service, participant intellectual property).

- Did you consult with anyone, consider other research projects, or refer to literature, policies, or guidelines regarding this issue? Please explain.
- Do you consider your research to fall under Fair Use?

Additional issues

7. Are there issues or challenges that arose during your example that I haven’t asked you about?

8. Who else should I interview? I am trying to reach big social researchers, qualitative researchers who have published or reused data, and data curators who have worked with qualitative or big social data.

Appendix 5. Data curators interview guide

Informed consent

We will review and sign the consent form.

Project Overview

Project Title

Connecting communities of practice: Using strategies from qualitative data curation to support big social research

Research summary

Big social data (such as social media and blogs) and archived qualitative data (such as interview transcripts, field notebooks, and diaries) are similar, but their respective communities of practice are under-connected. Research with both types of data repurpose existing social data to advance discoveries in social science. However, despite these similarities, big social research has not yet been widely framed as a form of qualitative data reuse, and qualitative data reuse has only begun to be discussed through a big social data lens. Qualitative data reuse is a more established practice, and therefore has more developed data curation strategies to support data sharing. My research investigates how data curation practices from each of these communities can inform the other for mutual benefit. The research will use interviews of qualitative researchers, big social data researchers, and data curators to gain insights into different community approaches to research and data sharing.

Research background

This research asks: how can data curators best handle qualitative and big social data to support ethical, epistemological, and legal data sharing practices?

My review of the literature revealed that there are six key issues that pose challenges for both groups. During the interview, I will ask questions about your personal experience in each of these six topic areas, plus introductory and wrap up questions. The interview will take 60-75 minutes.

Thank you for taking the time to speak with me! Your interview will help to improve data curation practices across disciplines.

Interview questions

We'll start recording the interview now.

Introductory question

Tell me about the types of data you usually curate and what your interests are regarding data curation.

Identifying a specific example

Please describe a recent time when you:

- curated qualitative data for sharing;
- curated big social data for sharing;
- advised or collaborated with big social researchers on data collection and/or analysis; or
- observed firsthand someone else doing one of the above.

If you have worked with both qualitative data and big social data, please identify two examples—one for each type of data.

Was this example part of a grant-funded project that required specific treatment of the data?
Was there a data management plan?

If you supported publication for any of the data in your example:

- Is the data published in a repository? Which one?
- What are the plans for storing, retaining, and deleting data in the future?
- Who has access to the data?

Context

Qualitative data context

Qualitative research is a process that may include deep and prolonged contact and connection with research subjects, attempting to understand subjects within their own context (Miles et al., 2020). Qualitative data are therefore highly context-dependent. As Hinds et al. write, “context is a source of data, meaning, and understanding… Ignoring context, underusing it, or not recognizing one's own context-driven perspective will result in incomplete or missed meaning and a misunderstanding of human phenomena” (1992, p. 72).

Big social data context

Halavais (2015) suggests that “when we collect data from [social media] platforms (just as when we collected data in traditional spaces), context matters.” However, the context of a social media post may be absent or difficult to understand. Social media posts are by nature short pieces of text, taken from a larger context of personal and

public life. This out-of-context effect is only compounded when data are amassed at a large scale.

1. During your example, what challenges did you encounter (if any) when trying to capture the context in which the data was collected?

- What strategies did you use to document context for future users (e.g. metadata, documentation, linking related datasets)?
    - Please describe in detail how these strategies helped you.

1a. (If applicable) What similarities and differences do you see between data curation strategies that address context issues for qualitative data and big social data?

Data quality

2. During your example, what challenges did you encounter (if any) when trying to document data quality (e.g., missing data, bots, bias, quality of method)?

- what strategies did you use to communicate, describe, or clarify data quality issues in your example?
    - Please describe in detail how those strategies helped you.
- What factors helped you and the researcher better communicate/document data quality issues for future users? What factors prevented you from communicating data quality?

2a. (If applicable) What similarities and differences do you see between data curation strategies that address quality issues for qualitative data and big social data?

Data comparability

3. During your example, what challenges did you encounter (if any) relating to comparability or interoperability of your dataset? (e.g. missing data, different research questions, different methods, metadata interoperability)?

- What strategies did you use to address these challenges?

3a. (If applicable) What similarities or differences did you see regarding data comparability for qualitative data and big social data?

Informed consent

4. In your example, what challenges did you encounter (if any) relating to informed consent for participants, particularly consent for future use of the data?

- What strategies did you use to address these challenges?

4a. (If applicable) What similarities or differences did you see regarding informed consent for qualitative data and big social data?

Privacy and confidentiality

5. During your example, what challenges did you encounter (if any) relating to privacy for the people represented in the data?

- What strategies did you use to address these challenges (e.g. restricted access, de-identification, publishing limited metadata)?

5a. (If applicable) What similarities or differences did you see regarding privacy for qualitative data and big social data?

- Did you encounter different challenges when protecting privacy for qualitative data or big social data?

Intellectual property

6. During your example, what challenges did you encounter (if any) regarding intellectual property concerns of archiving/publishing data (e.g. for qualitative data: participant IP; e.g. for big social data: social media terms of service)?

- What strategies did you use to address these IP concerns and issues?
- What do you feel are your responsibilities as a data curator regarding intellectual property?
- Did you consider data sharing to fall under Fair Use?

6a. (If applicable) What similarities or differences did you see regarding intellectual property for qualitative data and big social data?

Additional issues

7. Are there issues or challenges that arose during your example that I haven't asked you about?

8. Who else should I interview? I'm looking for big social researchers, qualitative researchers who have published or reused data, and data curators who have worked with qualitative or big social data.

Appendix 6. Interview dates and lengths

Interview code	Date	Length (minutes)
QR01	3/11/2021	54:02
BSR01	3/31/2021	34:24
DC01	4/5/2021	48:52
DC02	4/8/2021	68:49
DC03	4/12/2021	43:00
DC04	4/15/2021	65:55
DC05	4/21/2021	40:43
DC06	4/27/2021	60:13
BSR02	4/29/2021	57:20
QR02	4/30/2021	51:01
QR03	5/6/2021	67:17
DC07	5/17/2021	70:39
DC08	5/17/2021	72:11
BSR03	5/19/2021	50:29
QR04	5/26/2021	51:00
DC09	5/28/2021	70:21
DC10	6/2/2021	54:15
BSR04	6/17/2021	58:21
BSR05	6/28/2021	50:40
QR05	7/15/2021	44:04
BSR06	7/29/2021	49:56
QR06	8/11/2021	57:52
QR07	8/25/2021	40:10
BSR07	8/30/2021	43:06
BSR08	9/2/2021	73:25
QR08	9/2/2021	55:04
BSR09	9/3/2021	35:13
QR09	9/7/2021	32:36
QR10	9/7/2021	42:37
BSR10	10/6/2021	47:06

Appendix 7. Invitation emails to participants

Big social and qualitative researchers

Subject: Research interview request

Dear <Name>

I hope this email finds you well.

I am a librarian at Montana State University and a doctoral candidate at Humboldt University in Berlin (HU Supervisor: Vivien Petras; External Supervisor: Michael Zimmer, Marquette University). I am conducting a research project that aims to understand how different research communities address data curation and data sharing. The results of the research will improve library and data repository practices.

I am interviewing researchers to collect data for the project. (See full summary of the project below this email.) I reviewed your <Year> article, <"Title">, and I believe that you will be able to provide valuable insights into big social research.

My request to you:
Would you be willing to join me for a 60-minute research interview?

If you're available and interested, I will be conducting interviews throughout the next couple of months (through October). You can select a time slot within the next several weeks that works for you using my bookings page <link>. The system will automatically schedule an appointment with Zoom information.

Thank you very much for considering!

Sincerely,
Sara

—
Sara Mannheimer (she/her)
Associate Professor, Data Librarian - Montana State University
Doctoral Candidate - Humboldt University of Berlin
https://saramannheimer.com

Full research summary

Big social data (such as social media and blogs) and archived qualitative data (such as interview transcripts, field notebooks, and diaries) are similar, but their respective communities of practice are under-connected. Research with both types of data repurpose existing social data to advance discoveries in social science. However, despite these similarities, big social research has not yet been widely framed as a form of qualitative data

reuse, and qualitative data reuse has only begun to be discussed through a big social data lens. Qualitative data reuse is a more established practice, and therefore has more developed data curation strategies to support data sharing. My research investigates how data curation practices from each of these communities can inform the other for mutual benefit. The research will use interviews of qualitative researchers, big social data researchers, and data curators to gain insights into different community approaches to research and data sharing.

Data curators

Subject: Research interview request

Dear <Name>,

I hope this email finds you well.

I am conducting a research project that aims to understand how different communities of practice use data curation to support ethical, legal, and epistemologically-sound big social research. I am conducting interviews to collect data for the project. (See full summary of the project below this email.)

My request to you:
Would you be willing to join me for a research interview? The interview will take 60 minutes. <Because of your job position/experience>, I believe you will have valuable knowledge and experience about curating qualitative and big social data.

Thank you very much for considering!

Sincerely,

Sara

—
Sara Mannheimer (she/her)
Associate Professor, Data Librarian
Montana State University
https://saramannheimer.com

Full research summary
Big social data (such as social media and blogs) and archived qualitative data (such as interview transcripts, field notebooks, and diaries) are similar, but their respective communities of practice are under-connected. Research with both types of data repurpose existing social data to advance discoveries in social science. However, despite these

similarities, big social research has not yet been widely framed as a form of qualitative data reuse, and qualitative data reuse has only begun to be discussed through a big social data lens. Qualitative data reuse is a more established practice, and therefore has more developed data curation strategies to support data sharing. My research investigates how data curation practices from each of these communities can inform the other for mutual benefit. The research will use interviews of qualitative researchers, big social data researchers, and data curators to gain insights into different community approaches to research and data sharing.

Appendix 8. Follow up emails to participants

Wonderful! Thank you for your generosity with your time during a hectic year.

To prepare for the interview, I'm asking respondents to identify a recent time when you:

<For data curators:
- curated qualitative data for sharing;
- curated big social data for sharing;
- advised or collaborated with big social researchers on data collection and/or analysis.>

<For big social researchers:
- collected big social data for research; or
- reused big social data that was shared by someone else; or
- prepared big social data for publication or sharing; or
- considered sharing your big social data, even if you ended up deciding against sharing.>

<For qualitative researchers:
- prepared your qualitative data for publication or sharing; or
- reused existing qualitative data yourself; or
- considered sharing your qualitative data, even if you ended up deciding against sharing.>

I have attached the full interview guide and IRB-approved consent form for your reference—feel free to review them ahead of the interview, but don't feel that you have to. We'll review and sign the consent form using DocuSign on the day of the interview.

Thanks again, and I look forward to talking with you,

Sara

—
Sara Mannheimer (she/her)
Associate Professor, Data Librarian
Montana State University
https://saramannheimer.com

Appendix 9. Thank you email to participants

Subject: Thank you!

Dear <Name>,

Thank you so much for taking the time to talk with me. Your experiences and insights will be a key addition to my study, and I hope the results will support new knowledge in data curation for responsible qualitative data reuse and big social research.

I truly appreciate your thoughtfulness and your time.

Sincerely,

Sara

—
Sara Mannheimer (she/her)
Associate Professor, Data Librarian
Montana State University
https://saramannheimer.com

Appendix 10. Initial codebook

Codebook abbreviations

API - application programming interface
bsd - big social data
bsr - big social research
CITI - Collaborative Institutional Training Initiative
IP - intellectual property
IRB - institutional review board
QDR - Qualitative Data Repository
QDAS - qualitative data analysis systems
qual - qualitative
Refs - references
SEO - search engine optimization

Initial Code	Files	Refs
benefit of bsr	1	1
benefit of bsr - requires fewer resources	1	1
bsd collection - iterative process	1	1
bsd collection log like field notes - documentation of process	1	1
comparability - complexity of qual data	2	2
comparability - contextual documentation supports comparability	1	1
comparability - data dictionary	1	1
comparability - didn't think about it	2	2
comparability - different languages	1	1
comparability - documentation to support interoperability	2	2
comparability - expanding research	2	2
comparability - file formats	7	7
comparability - helpful, auto-generated reports from Social Feed Manager	1	1
comparability - include summary tables, codebooks	1	1
comparability - interoperability between QDAS	1	1
comparability - interoperability within a project	1	1
comparability - lack of standards	2	2
comparability - manual matching of different datasets	2	2
comparability - matching social media users using names is difficult	1	1
comparability - metadata standards	3	3
comparability - more data = stronger conclusions	8	8
comparability - page limits of journals - hard to explain complex different datasets	1	1
comparability - providing training and code with published data—to help	1	1

future users read and analyze data		
comparability - publishing codebooks	1	1
comparability - uneven levels of description for archived bsd	1	1
comparability - using manual search to find historical tweets	1	1
consent - adding a discussion of ethical practice early in the paper	1	1
consent - altering timelines and changing data requires IRB approval	1	1
consent - API as a tool doesn't encourage interaction with users	1	1
consent - asking permission for direct quotes	5	6
consent - be careful with direct quotes	3	3
consent - big social data - trends in IRB requirements	2	2
consent - biggest challenge of the six	1	1
consent - bsd archiving for historical record	3	4
consent - bsr and data reuse not seen as human subjects research	5	5
consent - clear consent is rare	1	1
consent - collecting tweets can feel invasive	1	1
consent - concern that consent to data sharing would suppress participation	2	2
consent - consent form language affects what can be shared	1	1
consent - continual consent	1	1
consent - curation workflows depending on consent procedures	1	1
consent - decision tree to see if data is shareable	1	1
consent - definition of data sharing	1	1
consent - deidentifying social media posts	2	2
consent - don't know what future uses might be	2	3
consent - education on how to get consent for data sharing	1	1
consent - experiment affected users standing in social media community	1	1
consent - extremely uncommon for big social data	2	2
consent - for bsd, consent may need to come from whole community	1	1
consent - forms explicitly say data won't be shared	1	1
consent - going through IRB supports responsible research	4	4
consent - harm analysis	2	3
consent - if consent is unclear, a repository can restrict access	2	2
consent - if forms don't talk about data sharing at all, can open the door to deidentified sharing	1	1
consent - if users knew they were part of an experiment, could skew results	1	1
consent - impractical with bsr	1	1
consent - IRB	3	3
consent - IRB classifies bsr and data reuse as exempt	10	10
consent - IRB not necessary for bsd	1	1

consent - IRB template not sufficient	1	1
consent - is reuse aligned with intent of original study	1	1
consent - learning strategies from advisors, other researchers	1	1
consent - little guidance	2	3
consent - more important with vulnerable populations	2	2
consent - no consent for bsr	1	1
consent - one of the biggest issues for curation	1	1
consent - participant review and redaction of transcripts	4	4
consent - participants may not understand nuances of consent form	3	4
consent - planning ahead for sharing	3	3
consent - presentation and Q&A to explain consent procedures	1	1
consent - public vs. private	8	9
consent - quality and content of consent forms vary	3	3
consent - quoting tweets didn't feel right	1	1
consent - re-consent to support data sharing	1	1
consent - reconsent	1	1
consent - reconsent is often impossible or impractical	1	1
consent - renewed consent for longitudinal studies	1	1
consent - repository requires proof IRB review or exemption	1	1
consent - repository terms of use	2	3
consent - research ethics education	2	2
consent - research ethics literature	2	2
consent - researchers don't think to ask for consent for data sharing	1	1
consent - sensitivity of data	3	3
consent - social media terms of service include consent	5	5
consent - social media users would be okay with data being used for good ends	1	1
consent - some bsr platforms are more public than others	2	3
consent - some consent forms say nothing will be shared	1	1
consent - some documentation is better than none - even if consent issues constrain sharing	1	1
consent - tiered consent	3	3
consent - tiered consent can lead to missing data	1	1
consent - tiered consent meant less data available to share	1	1
consent - to data reuse	6	8
consent - tools to support consent in bsr	1	1
consent - trust in data repository procedures	1	1
consent - trust in data reusers	1	1

context - big social data - interface provides context	1	1
context - bsr - researcher data collection vs. platform newsfeed	1	1
context - can't ask follow up questions of bsd participants	1	1
context - collaborate with original authors	1	1
context - collecting bsd in the moment doesn't account for future interaction	1	1
context - consult with IRB or research compliance office	1	1
context - consulting with data curators	1	1
context - creating a standard for minimum viable metadata	1	1
context - description, metadata, documentation to support context	10	11
context - description, metadata, documentation to support context\context - dataset metadata in addition to article	2	2
context - description, metadata, documentation to support context\context - document how data were collected	2	2
context - description, metadata, documentation to support context\context - document info about population	1	1
context - description, metadata, documentation to support context\context - much qualitative metadata is unstructured	1	1
context - description, metadata, documentation to support context\context - readme files	1	1
context - did not document these issues	1	1
context - different disciplinary expectations	3	3
context - different research methods provide different contextual info	3	3
context - filling gaps in hashtag data by retroactively collecting timeline data	1	1
context - focusing bsr on a hashtag or space, not individual, supports context	1	1
context - full context of qualitative research is difficult to document	3	3
context - good documentation is time consuming	5	6
context - identifying specific users to collect bsd	1	1
context - in tension with privacy	10	12
context - in which data was posted or collected vs in which it will be used	3	3
context - including related materials with data	4	6
context - key part of why you do qualitative research	1	1
context - key to understanding reused data	1	1
context - linked to related research and data	1	1
context - longitudinal qualitative studies	1	1
context - look to existing literature for guidance	1	1
context - may be difficult to ascertain with bsr	6	7
context - misinterpretation may be inevitable	3	3
context - more data supports context and quality	1	1

context - peer review feedback	1	1
context - providing enough, but not too much information	2	2
context - publishing more data to support context	1	1
context - publishing reproducible code	2	2
context - purpose of data collection	1	1
context - qualitative analysis of social media helps understand context	1	1
context - qualitative research reporting standards for articles	1	1
context - reading text or API results vs on the platform	2	2
context - representativeness of data	3	5
context - research design	1	1
context - researchers will inevitably bring new context	1	1
context - researchers, reusers, curators have different backgrounds	4	5
context - reuse own data	1	1
context - reviewing similar papers for guidance	1	1
context - social media users are worldwide, but geotags are rare	1	2
context - software used to collect bsd	1	1
context - some data have more inherent context	3	3
context - some data is better than none, regardless of how well-documented	1	1
context - tagging and OCR for searchability	1	1
context - trust in data creators	2	2
context - using platform demographics to provide more context	1	1
context - when reusing data, using the same pseudonyms to provide continuity from article to article	1	1
curation - benefits of sharing data	4	5
curation - benefits of sharing data - less burden on respondents	1	1
curation - benefits of sharing data - reduces cost for secondary researchers	1	1
curation - can be difficult to reach PIs and data depositors	1	1
curation - codebooks and storytelling for transparency	1	1
curation - collaborating with curators and repositories	2	2
curation - concern about being scooped	1	1
curation - concern about cost	1	1
curation - connection btw bsd and web archives	1	1
curation - consent form review	1	1
curation - considering reusers needs when publishing data	1	1
curation - content warnings	1	1
curation - data authorship may be different from article	1	1
curation - data citation practices	1	1
curation - data reuse is rare	1	1

curation - for transparency	3	3
curation - good enough metadata is sometimes as good as it gets	2	2
curation - helping with deidentification	1	1
curation - highlighting benefits of FAIR data sharing	1	1
curation - iterative process	1	1
curation - large size data difficult to publish	1	1
curation - levels of curation	1	1
curation - metadata and documentation	2	2
curation - more controversial papers require more rigorous reproducibility strategies	1	1
curation - more efficient for repository to provide analysis rather than full dataset	1	1
curation - new ways of indexing and providing access to bsd	1	1
curation - our job is only data, not epistemology	1	1
curation - partnership with IRB	1	2
curation - planning for data sharing makes it less of a hurdle	3	4
curation - platform terms of service	2	2
curation - publish tweetIDs, rather than full data	1	1
curation - QDR important to supporting qual data sharing	1	1
curation - questionnaires, codebooks publicly available, even for restricted data	1	1
curation - repo and library resources for deidentification guidance	1	1
curation - repository providing analytical output of data rather than full dataset	1	1
curation - repository quality standards	1	1
curation - restrict data linkages to support privacy	1	1
curation - sharing a subset of bsd because full dataset is too large	1	1
curation - still possible for data to be misunderstood	1	1
curation - strategies to ease burden on researchers	1	1
curation - technical requirements of bsd	1	1
curation - time-consuming	9	10
data retention	1	1
data sharing - contact info for original researcher	1	1
data sharing - disciplinary values and norms	3	3
data sharing - github not good for big datasets	1	1
data sharing - individual sharing with students	1	1
data sharing - librarian and curator support	1	1
data sharing - not necessary for bsd bc publicly available	1	1

data sharing - repository SEO, data findability	2	2
data sharing - restrictions with purchased data	1	1
data sharing - social media terms of service restrictions	1	1
data sharing as opportunity to add value to archival data	1	1
data sharing honors the gift respondents have given of their time and experience	1	1
data sharing requirements	2	2
data sharing too risky	1	1
disciplinary - bsr collaboration with social scientists	3	3
disciplinary - different disciplines have different research practices and standards	5	5
disciplinary - difficult to find computational expertise plus social science expertise	1	1
disciplinary - divide between computer science and bsr	1	1
disciplinary differences between qualitative and big social researchers	1	1
disciplinary ideas - qualitative researcher as instrument	1	1
epistemological discussions with colleagues and collaborators	5	5
ethical codes for different professions and disciplines	1	1
ethical discussions with colleagues and collaborators	8	13
ethical guidelines for bsr	2	2
ethical review depends on where the data will be available	1	1
ethics - idea of respectful reuse	1	1
ethics-related literature	2	2
evolving bsr availability	1	1
evolving ideas about data sharing	6	9
evolving representativeness of different social media platforms	2	2
evolving research ethics and values	3	3
evolving research methods	2	3
evolving usage of social media platforms	2	2
field notes and transcripts donated at end of researcher's career	1	1
human subject vs historical data	2	2
IP - ask permission to republish	1	1
IP - bending social media terms of service	3	4
IP - checking with participants and organizations involved	1	1
IP - citation to support IP	2	2
IP - community-driven research and data governance	2	2
IP - consult with legal	1	1
IP - Creative Commons license	2	3

IP - creative works like memes posted on social media	1	1
IP - data belongs to researcher's institution	3	3
IP - data citation	1	2
IP - data licensing	3	3
IP - data sovereignty and ownership	3	4
IP - developer terms of service	2	2
IP - evolving terms of service can impact reproducibility	2	2
IP - fair use	2	2
IP - following social media terms of service	7	7
IP - generally doesn't come up	2	2
IP - historical documents from commercial databases	1	1
IP - if curators add value to data, changes IP	1	1
IP - if not patentable info, not as important	1	1
IP - if the researcher leaves the institution, may not be able to publish the data anymore	1	1
IP - lack of clarity about IP laws	2	2
IP - legal aspects not clear to researchers	2	2
IP - legal gray areas	1	1
IP - legal ramifications of breaking terms of service	3	3
IP - legal restrictions for big social data collection	1	1
IP - of social media users	2	2
IP - privacy of participants more important than IP	1	1
IP - publish key findings from data before publishing data	1	1
IP - repository terms of use	3	3
IP - researchers did not read platform terms of services	1	1
IP - reused copyrighted materials	2	2
IP - scraping databases	1	1
IP - sharing a subset of data	1	1
IP - terms of service change over time	3	3
IP - terms of service not ethical rules	2	2
IP - use care when using copyrighted materials	2	3
IP - using or buying existing big social datasets	1	1
IP- different policies in different countries	1	1
method - proactive data collection on a hashtag	1	1
methodology - asking the right research questions for big social data	4	4
methodology - computational and manual	1	1
methodology - qualitative coding for bsd	1	1
methods - alternative methods to support consent and privacy for bsr	1	1

more research on Twitter, which has easier data collection	1	1
not much is known about qualitative data reusers	1	1
not using data if it wasn't collected ethically	1	1
outdated viewpoints or methods	1	2
partial data sharing	1	1
participant involvement in decision-making	1	1
power dynamics of research	3	3
preservation of links in big social data as preservation of context	1	1
privacy - aggregating data	1	1
privacy - assembling a lot of big data can threaten privacy	2	2
privacy - avoiding research that could be viewed as surveillance	1	1
privacy - bsd intended for academic use - like archival collection	1	1
privacy - bsd that wasn't relevant to research but was collected	1	1
privacy - care to make sure quotes aren't identifiable	2	2
privacy - challenges of deidentification	9	16
privacy - check back with participants	2	2
privacy - collecting public figures bsd vs private individuals	5	6
privacy - considering potential harms	3	4
privacy - creating fake tweets to demonstrate the algorithm	1	1
privacy - customizing terms of restricted access	4	4
privacy - data collection methods to support privacy	2	2
privacy - data curator is final authority	1	1
privacy - data security	5	5
privacy - datasets provided by social media companies	2	2
privacy - deidentification	2	2
privacy - deidentification doesn't harm analysis	1	1
privacy - deidentification strategies	2	3
privacy - deidentifying according to the consent form language	1	1
privacy - deletion requests	1	1
privacy - depositor-approved access - not ideal for long-term access	1	1
privacy - different policies in different countries	1	1
privacy - difficult to deidentify video	1	1
privacy - disclosure risk review	3	3
privacy - does the responsibility fall on curators or researchers	1	1
privacy - don't ask questions that might put participants in danger	1	1
privacy - embargo period	1	1
privacy - experience with sensitive data informs bsr	1	1

privacy - hashtags as a public space	2	2
privacy - how to handle deleted posts in dataset	1	2
privacy - importance of deidentification	1	1
privacy - IRB regulates confidentiality	1	1
privacy - keep annotations private, even if data is public	1	1
privacy - keeping data private reduces reproducibility	1	1
privacy - lack of agreement about social media data sharing	1	1
privacy - not collecting protected tweets	1	1
privacy - not everything needs to be deidentified	2	2
privacy - not seen as an issue for publicly available data	1	1
privacy - open access data more important to deidentify	1	1
privacy - participant expectations	10	13
privacy - potential harms of using data	1	1
privacy - pseudonyms	1	1
privacy - remove tweets that are too unique	1	1
privacy - repository data security	1	1
privacy - research design	1	1
privacy - research ethics education	2	2
privacy - restricted access	9	11
privacy - restricted access - ethics training required to access	1	1
privacy - restricted access - physical location	1	1
privacy - restricted access is effective	1	1
privacy - restricted access is safer than deidentification	1	1
privacy - retweets	1	1
privacy - reuse restrictions - CITI training to use the data	1	1
privacy - reuse restrictions - repository terms of service	2	2
privacy - reuse restrictions - secondary IRB approval	2	2
privacy - reuse restrictions - submit research plan to use the data	1	1
privacy - sensitivity of data	11	11
privacy - sharing codes but not transcripts	1	1
privacy - staffing and staff training	1	1
privacy - try to collect and save as little data as possible	1	1
privacy - twitter users who are no longer living	1	1
privacy - user expectations not an issue	1	1
publishing Tweet IDs allows people to delete or protect account, thus opt out	1	1
purposeful identification of research participants	2	3
qual data consent - if no consent, must be completely deidentified	1	1

qual data reuse - not well documented how archived data is used	1	1
qualitative data not the same as big social, but should also be shared	1	1
qualitative data reuse is rare	2	2
qualitative research - looking at field notes for patterns and gaps - similar to social media	1	1
qualitative research - not generalizable	1	1
qualitative researcher relationship with participants	2	2
qualitative researchers more thoughtful about human subjects considerations	2	2
quality - automated deletion of spam and bots	4	4
quality - Big social data are so big, just check if it loads, if code works, any major issues	1	1
quality - bsr data loss over time	2	2
quality - clearly document completeness	4	5
quality - combining datasets to support quality	2	2
quality - curation workflows can be arduous	1	1
quality - curator review	1	1
quality - curators can only assess quality of documentation, not method or bias	1	1
quality - curators have more responsibility to carefully review qualitative data	2	2
quality - data cleaning requires judgment calls	1	1
quality - data completeness	1	1
quality - data repository collects data in tandem with researchers	1	1
quality - demographics	1	1
quality - description, metadata, documentation support data quality	14	15
quality - differences in understanding transcripts or videos	1	1
quality - documenting potential bias	2	2
quality - documenting sampling technique	3	3
quality - incomplete dataset is shared	1	1
quality - increasing documentation could help	1	1
quality - look to existing literature for guidance	2	2
quality - low video quality reduces nonverbal cues	1	1
quality - misread unicode characters	2	2
quality - missing bsr data could go unnoticed	1	1
quality - missing data	3	3
quality - qual data reuse - new questions may not be answered as in-depth	1	1
quality - rehydrated tweetIDs can result in missing data	2	2
quality - repository quality control process	1	1
quality - representativeness - proportion of tweets available	3	4

quality - researcher bias	2	2
quality - shadow banned users	1	1
quality - size of data affects how well you can document quality	2	2
quality - sometimes bots are relevant	3	4
quality - spam in bsr	1	1
quality - testing different data samples against each other	1	1
quality - transcript quality	2	3
quality - truncated tweet IDs in Excel	1	1
quality - trust in data creator	3	4
quality - unclear and changing social media platform and API practices	2	2
researcher reluctance to share data	1	1
reusing bsr	1	1
risk-benefit analysis	11	18
risk-benefit of breaking terms of service - these protect sm company, not users	2	2
risk-benefit of openness vs privacy	7	8
scope - challenges of getting enough funding to do big, consent-based studies on social media	1	1
scope of conclusions	1	3
sharing data - fresh eyes bring new meaning	1	1
sharing reddit data	1	1
synthesis - data quality - similar issues with big social and qual data	1	1
synthesis - different disciplines approach the idea of human subjects differently	3	3
synthesis - people more hesitant to share qual data than social media posts	1	1
synthesis - qual researchers looking to big social researchers for innovative ideas	1	1
synthesis - scaling up qual research - not common but maybe growing	1	1
tension between benefit of data sharing and risk of harm	2	2
understanding participant communities	1	2
usefulness of sharing data	1	1
using different social media platforms to reach different demographic populations	2	2
Value decreases as data gets older	2	2
value despite data inaccuracies	1	1
what is an acceptable level of risk	1	1

Appendix 11. Final codebook

Codebook abbreviations

API - Application programming interface
bsd - Big social data
bsr - Big social research
IP - Intellectual property
IRB - Institutional review board
Refs - References
qual - Qualitative

Name	Files	Refs
comparability - complexity of qual data	2	2
comparability - documentation and metadata	9	11
comparability - interoperability - formats, metadata, language, etc	11	14
comparability - matching of different datasets	5	6
comparability - more data = stronger conclusions	10	10
consent - as it applies to whole communities	2	3
consent - biggest challenge of the six	2	2
consent - bsd archiving for historical record or unknown future use	4	8
consent - concern that consent issues would affect participation	3	3
consent - consent language and procedures	15	28
consent - don't know what future uses might be	5	6
consent - harm analysis	3	5
consent - human subject vs historical data	2	2
consent - IRB	22	33
consent - little guidance	3	4
consent - participant review and redaction of transcripts	4	4
consent - participants may not understand nuances of consent form	4	5
consent - public vs. private	8	9
consent - reconsent	2	4
consent - repository terms of use	2	3
consent - research ethics education and literature	4	6
consent - sensitivity of data	5	5
consent - social media terms of service include consent	6	6
consent - some bsr platforms are more public than others	3	4
consent - taking care with direct quotes	7	9
consent - tiered consent	3	5
consent - uncommon for big social data	5	6

context - big social data - interface and features provides context	6	8
context - can't ask follow up questions of bsd or qual reuse participants	3	3
context - description, metadata, documentation to support context	13	21
context - different disciplinary expectations	3	3
context - different research design and methods provide different contextual info	4	4
context - filling gaps in hashtag data by retroactively collecting timeline data	2	2
context - good documentation is time consuming	7	9
context - in tension with privacy	10	12
context - in which data was posted or collected vs in which it will be used	4	4
context - including related materials with data	9	13
context - involve original authors for reanalysis	2	2
context - key to understanding reused data	3	3
context - look to existing literature for guidance	3	3
context - may be difficult to ascertain with bsr	7	8
context - misinterpretation may be inevitable	3	3
context - providing enough, but not too much information	3	3
context - representativeness of data	5	7
context - reproducibility	4	5
context - researchers, reusers, curators have different backgrounds	4	5
context - some data have more inherent context	4	5
context - trust in data creators	2	2
curation - collaborating with curators and repositories	7	8
curation - considering reusers needs when publishing data	2	2
curation - cost and time	10	11
curation - data sharing requirements	2	2
curation - for transparency	4	6
curation - good enough metadata is sometimes as good as it gets	2	2
curation - planning for data sharing makes it less of a hurdle	5	6
curation - qualitative data reuse is rare and hard to track	3	3
curation - repository SEO, data findability	2	2
curation - researcher reluctance to share data	3	3
curation - technical requirements of bsd and data reuse	4	8
curation - value of bsr and qual data sharing	10	18
domain differences - bsr collaboration with social scientists	4	4
domain differences - data sharing values and norms	12	19
domain differences - skills, training, and background	8	8
domain differences - research practices and standards	9	12

IP - checking with participants and organizations involved	3	3
IP - citation to support IP	5	7
IP - data licensing	6	7
IP - data sovereignty and ownership	7	11
IP - fair use	2	2
IP - lack of clarity about IP laws	5	5
IP - of social media users	2	2
IP - platform or data provider terms of service	13	28
IP - purchasing or using commercially-available data	8	10
IP - repository terms of use	3	3
IP - sharing a subset of data	2	2
privacy - assembling a lot of big data can threaten privacy	2	3
privacy - care to make sure quotes aren't identifiable	3	3
privacy - check back with participants	2	2
privacy - considering potential harms	10	14
privacy - data security	6	6
privacy - datasets provided by social media companies	2	2
privacy - deidentification	18	36
privacy - deletion requests	3	5
privacy - partial sharing to support privacy	2	3
privacy - participant expectations	10	13
privacy - research design	8	14
privacy - research ethics education and training	3	3
privacy - restricted access	11	25
privacy - sensitivity of data	11	11
quality - bsr data loss over time	2	2
quality - combining datasets to support quality	2	2
quality - curator review	4	6
quality - data completeness	10	15
quality - description, metadata, documentation support data quality	18	23
quality - issues with large-scale and automated collection	7	9
quality - look to existing literature for guidance	2	2
quality - representativeness of data	5	6
quality - researcher bias	3	3
quality - spam and bots	6	10
quality - trust in data creator	3	4
quality - understanding transcripts or videos as opposed to in person	3	5

strategies for responsible practice - appropriate research questions and scope	5	8
strategies for responsible practice - considering power dynamics of research	4	4
strategies for responsible practice - discussions with colleagues and collaborators	13	19
strategies for responsible practice - ethical guidelines for bsr	3	4
strategies for responsible practice - risk-benefit analysis	17	32

Appendix 12. Memos

Big social researchers

BSR01

Shares social media data, has more concerns about reproducibility than privacy
More motivated to get research results, make sure to follow terms of service
Spoke with Reddit moderators
But didn't consider ethical challenges related to the idea of BSR as human subjects research
From CS, publishes in management literature as well, aware of disciplinary differences in how data collection and BSR are addressed

BSR02

big social data is controlled by companies/organizations—database changes, data format, data sampling/data provided, terms of service, etc.
Idea of participants wanting credit for their contributions, depending on the community (e.g. on Wikipedia there is a value of openness and credit)

BSR03

Metadata quality, shortcomings of "authority records"
I was especially struck by the idea of specifically designing research questions in a way that takes into account ethical considerations. This can help guide big social researchers toward ethical research questions. However, it may limit the research questions that can responsibly be asked of research data. How might researchers ask questions that may be more sensitive, while still maintaining ethical standards?
only try to measure things that you think will require ethically sound methods

BSR04

training and values of different academic research communities
IRB not providing guidance, so looking to colleagues and literature for guidance
tension between openness/transparency and privacy
privacy a way to handle issue of consent

BSR05

weighing risk with reward
will the results of the research be important enough to justify risk to participants? esp with sensitive data, or identified data like the panel with voter records matched with Twitter
Openness vs privacy.

- How do researchers conduct these risk-benefit analyses? Have they been trained to do so?

- This is a classic strategy for understanding ethically-relevant harms, but most researchers appear to conduct these analyses informally.
- Analyses are done by talking to colleagues, reading relevant literature, and thinking about potential harms to participants (+ harms to reputation or other professional consequences).

Twitter ToS vs. better quality data
Reading, Conversations with collaborators a main way that he considered these issues

BSR06
Knowledge and thoughtfulness about responsible research is growing all the time - would have a better understanding now than in 2017

Collaboration with social scientists - as mentor, not coauthor

enough TweetIDs being published about natural disasters that he could reuse a lot of data, also collect his own

knowledge about who uses twitter

BSR07
- Different social media platforms have different context expectations for data. E.g. pinterest pins depend on the context of the board and related links, but are often taken out of context from the person who is pinning.
    - different social media platforms have different expectations of privacy - Pinterest is by nature less private—pins are being repinned all the time
    - However, most users wouldn't expect their pins to be used for research purposes—outside the context of their original intent/purpose of the pin
    - "So there's some concern there [about privacy], but I don't know, I feel like it's outweighed by the fact that, you know, we're trying to document something that might be harmful and trying to help public health professionals. So yeah, I feel like on balance, it's an acceptable practice."
- field of Journalism considers consent differently from qualitative researchers - not studying people, studying *content*
    - this also contributed to them not feeling they needed IRB approval
- Terms of service/fair use/IP confusion—didn't know how to think about it, just assumed it was okay, since they weren't publishing the full dataset

BSR08
protecting privacy
convenience of secondary datasets
pace of academic research is slow, but social media landscape changes quickly

BSR09

- because he bought the data from Twitter, not allowed to share.
- Some research questions work better with social media - e.g. weather disaster events
- used filters to filter tweets for correct context, but hard because limited metadata—not many geotags, developed a method to extract location from content of tweets
- tools to filter out bots
- didn't think about consent bc of Twitter terms of service - the team bought twitter data from the company, so felt fine using it without specific consent from users.
- also went through IRB, but felt if these larger entities okayed the research, they were okay without specific consent from users
- user privacy more important - protect identity of individuals - username, user ID, pictured in photos or named in tweets, took excerpts from tweets rather than quoting the full tweet
- had seen issues arise in previous social media papers (or in response to publication of previous papers), and responded to those issues

BSR10

- strategies for inferring context - profile data, hashtags, etc
- macroscopic level - by analyzing (topic modeling) more data, context will emerge.
- self regulatory behaviors in a social network - low quality bots or fake news spike, but then peter out quickly as influential members of the network
- NSF data sharing requirement - shared 1% of total tweet IDs (on project page, not repository) for convenience
- social media terms of service change often

Qualitative researchers

QR01

had a data curator on the research team who helped think through responsible sharing practices from the beginning of the research design
very thoughtful about data sharing, spent a lot of time working with QDR to make sure as much could be published safely as possible.

QR02

people have contacted them directly - just people who had read what they'd published
graduate students at the top of the pile get access to your data
data is transmitted in a "pseudo kinship" relationship
passed down to grad students after you die

data might get “stale” - useful to historians right to anthropologists

QR03
Committed to data sharing - has experienced others wanting to use her data
interested in the benefits of data sharing - to communities and to researchers
idea that respondents want to be identified - want their stories to be shared

QR04
only do research on data when you think you can ethically do so
scope conclusions as appropriate
qual researchers look to computer science to see how they can implement strategies that they use in CS to scale up

ethnographic reports as historical records rather than data per se - do your life’s work, then donate to libraries

QR05
Couldn’t find standardized protocols for deidentification, so created own protocol
Including a readme and links to related articles
The study included institutional data from their university that they couldn’t publish along with their research data

QR06
Had thought through curation with QDR people, but the terminology and ideas around data sharing were still pretty unfamiliar to them. Motivation to share data was funder mandate.
Context: team may think something is obvious, but maybe it’s not obvious to somebody who hasn't worked on this project for years.
how much of the consent form do participants really understand?
decided to deidentify videos (blur faces) even though participants had consented to having their faces in the videos.

QR07
training in qualitative research is helpful - if you don’t have it you have to cobble stuff together.
we need a good resource about deidentification of qualitative data. What are best practices? How can we make the data useful (maintain context), while not compromising privacy of participants?

QR08

Talked about how extensive the explanation of context should be, and how to balance contextual information with privacy/deidentification
Used quotes mined from research papers, so this is secondhand information already and the context is once removed.
Data from a blog where people post about diabetes—they considered it to be public, but they were also concerned with human participants.

QR09
reviewed the literature about how to do qual secondary analysis
research team knew and trusted each other
acted ethically according to "their own standards"

QR10
Secondary analysis was prompted by insights in the first analysis—analyzed own data.
Discussion within the research team about informed consent, but decided reconsent wasn't necessary
Strategies for reducing harm for participants—removing quotes critical of their workplace

Data curators

DC01
Considering how to reduce potential harms of responsible big social research—are there big social data that just shouldn't be collected and curated because it's too risky?
Considering how context can be communicated as part of the shared dataset, while still maintaining privacy. Finding balance.
Also discussed outreach and advocacy for library data services and data management.

DC02
Varied role and guidance provided by IRBS about data reuse.
Complexities of deidentification with qualitative data
Role of data curators to provide training and guidance on data sharing from the beginning of the research process.
Consent processes - tiered consent options
Restricted access/ access controls

DC03
qual curation is time consuming
consultation with colleagues

non-standardized metadata

DC04

Ensuring quality of transcripts - multiple transcribers and reviewers. Idea of data comparability in order to align with other studies. How to connect data to related studies. IP issues when datasets and data collection instruments are copyrighted.
Rarity of qualitative data sharing, even with a QDR membership

DC05

Data creators provided a document to the data curators explaining the deidentification procedures they had implemented so far

- what rules they used, what identifiers were masked, etc. Helped the curator follow those rules during their review.

Difficulty of responsiveness of data creators

- can the curators reach them with questions? Can be difficult to get responses, and therefore, curators may not reach out with questions.

Standard questions asked for privacy/disclosure risk review - how many subjects, vulnerable pops?, etc.

DC06

Embedded data curator in a research project. Felt most comfortable talking about support for data sharing - providing access. Curating longitudinal studies, making sure metadata was standardized through the years, across different students, lab members.

DC07

Ideas about archives/historical documents versus research objects

- collecting proactively - like oral histories

Generative/iterative activities - Learning more about the collection as you collect - new hashtags, new revisions - a thing in motion. parameters change over time.

- similar to how in qualitative research, you can learn more about your research question over time.

Reproducibility - because the social media dataset may change over time, can't really be used for reproducibility, but just transparency

- similar with qualitative data how you can't draw universal conclusions from the research, just for that specific population, and your specific context

DC08

context-dependent - different research questions, communities, datasets require different treatment
risk and benefit - just “learning something new” isn’t enough if there is a big risk to the community
idea of community versus individual consent and risk - informed consent from an individual doesn’t really matter if the community might be at risk
could be an argument for community focus groups or advisory boards for social media research projects, rather than individual consent
who is the curator/researcher responsible to?
Twitter? science? the community? weighing the responsibility to both and making decisions from there.
For social media data, repository providing analytical output of data rather than full dataset

DC09
flowchart for consent decision-making
moving into the big data space by talking about data sharing and reuse at computational social science conference
IRB connections
starting early in the process to support good data management
challenges of high quality curation when it may not be the PI’s priority
interoperability of qual data analysis systems

DC10
Data curator who was not trained as a librarian - thought that big social data was less identifiable, thought that more data could lead to broader conclusions
Less knowledgeable about library and information science disciplinary ideas

Appendix 13. Data availability: Transcripts and QDAS file

Deidentified transcripts and qualitative data analysis software (QDAS) files will be available in the Qualitative Data Repository (QDR) in late 2022. Transcripts are provided in Microsoft Word format (DOCX). QDAS files are provided in proprietary NVivo format (NVPX) as well as in an open format (QDPX) that is readable by any qualitative data analysis software.

Upon publication in QDR, the data will available with the following citation:

> Mannheimer, S. (2022). Data for: Connecting communities of practice: Data curation strategies for qualitative data reuse and big social research. Qualitative Data Repository. https://doi.org/10.5064/F6GWMU4O