Synthesis Lectures on
Information Concepts, Retrieval, and Services

SYNTHESIS
COLLECTION OF TECHNOLOGY

Sara Mannheimer

Scaling Up: How Data Curation Can Help Address Key Issues in Qualitative Data Reuse and Big Social Research

Springer

Synthesis Lectures on Information Concepts, Retrieval, and Services

Series Editor

Gary Marchionini, School of Information and Library Science, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

This series publishes short books on topics pertaining to information science and applications of technology to information discovery, production, distribution, and management. Potential topics include: data models, indexing theory and algorithms, classification, information architecture, information economics, privacy and identity, scholarly communication, bibliometrics and webometrics, personal information management, human information behavior, digital libraries, archives and preservation, cultural informatics, information retrieval evaluation, data fusion, relevance feedback, recommendation systems, question answering, natural language processing for retrieval, text summarization, multimedia retrieval, multilingual retrieval, and exploratory search.

Sara Mannheimer

Scaling Up: How Data Curation Can Help Address Key Issues in Qualitative Data Reuse and Big Social Research

Springer

Sara Mannheimer
Montana State University
Bozeman, MT, USA

ISSN 1947-945X ISSN 1947-9468 (electronic)
Synthesis Lectures on Information Concepts, Retrieval, and Services
ISBN 978-3-031-49221-1 ISBN 978-3-031-49222-8 (eBook)
https://doi.org/10.1007/978-3-031-49222-8

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

Acknowledgements

This book would not have been possible without my own communities of practice. Thank you to Scott Young, my partner and colleague; our discussions and mutual support enhance my life in general, and this book in particular. This book is based on my Ph.D. research, and I am deeply grateful for the guidance and support provided by my dissertation advisors, Vivien Petras and Michael Zimmer. Thank you to Kalpana Shankar for key advice on research methods; Eric Raile for reviewing my interview guides; Emily O'Brien for cleaning interview transcripts; and David Mannheimer for providing suggestions on writing style, clarity, and structure. Parts of Chap. 5 were previously published in the Journal of eScience Librarianship, and I am grateful to the JeSLIB peer reviewers whose feedback made those sections stronger. Thank you to Dessi Kirilova and the curators at Qualitative Data Repository for their seamless curation services when archiving the associated research data. Finally, I want to thank the thirty researchers and data curators who participated in research interviews. The knowledge, experience, and insights they shared are the heart of this book.

Contents

Introduction

1

> "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).

1.1 Background

The research community has recently seen increased interest in qualitative data archiving and reuse, in conjunction with shifts toward open science practices and engagement with new technologies (Corti et al. 2005; Glenna et al. 2019). There are many potential benefits of qualitative data reuse. For example, reusing qualitative data can increase efficiency, deepen research conclusions, and reduce the burden on research subjects by allowing new studies to be conducted without collecting new data. Qualitative data reuse can also potentially support 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). In the two

S. Mannheimer, *Scaling Up: How Data Curation Can Help Address Key Issues in Qualitative Data Reuse and Big Social Research*, Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-49222-8_1

decades since Mason issued this call, some researchers have aggregated qualitative data to produce new conclusions (Halford and Savage 2017; Winskell et al. 2018; Davidson 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, online forums, and posts and interactions on platforms like Facebook, Twitter, YouTube, and TikTok. These "big social data" (Manovich 2012) have been celebrated as unprecedented sources of data analytics, able to produce social insights by analyzing human behavior on a massive scale (Fan and Gordon 2014; Cappella 2017). Big social data are a form of qualitative data that have been published online by users themselves. When researchers analyze big social data, this could be seen as qualitative data reuse—that is, researchers are repurposing and recontextualizing big social data to answer research questions.

Using this similarity between qualitative data reuse and big social research as a starting point, this book investigates three communities of practice (Wenger et al. 2002) who are engaged with social research and social data:

- qualitative researchers who have shared or reused data
- big social researchers
- data curators.

Qualitative researchers who share or reuse data and big social researchers have similar goals—they aim to scale up and enhance social science research. But 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 research lens. These two communities of practice also have different backgrounds, training, and disciplinary values. Qualitative researchers tend to come from social science disciplines, and they tend to focus on using in-depth research methods to investigate social and behavioral phenomena. Big social researchers, on the other hand, 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.

Data curators as a profession are concerned with organizing, managing, and curating data, rather than building methodologies and drawing conclusions from those data. Therefore, data curators are uniquely positioned to build connections between qualitative researchers and big social researchers, based on the similarities of the data used by both types of researchers. In this book, I suggest that data curation strategies can be used to support and enhance responsible practice, and that data curators can act as facilitators and intermediaries between communities of practice.

1.2 Issues Raised by Qualitative Data Reuse and Big Social Research

This book is centered around six key epistemological, ethical, and legal issues that apply to qualitative data reuse, big social data research: context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. These six key issues are at the heart of this book, helping to structure interviews with researchers and curators, and functioning as scaffolding for data curators to build connections with researchers. Below, I provide brief summaries of each of the issues. These issues are addressed in more detail in Chap. 3 (as related to qualitative data reuse), Chap. 4 (as related to big social research), and Chap. 5 (comparing and contrasting issues for each type of research and synthesizing relevant data curation strategies for each issue).

1.2.1 Context

Both qualitative data reuse and big social research are context dependent. For qualitative data reuse, there is some concern that reused 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. For big social research, context is even more murky. Because automated data collection happens on a large scale, generally without interaction with the people who created the data, the context of big social data may be absent or difficult to understand.

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. Qualitative researchers who reuse data need to know that those data are high-quality and trustworthy—that the data have been collected using valid methods, that transcriptions are accurate, and that the data are complete. Big social researchers deal with the issue of data representativeness—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 data quality and trustworthiness are further complicated by the possibility of fake social media accounts and bots that may appear to be human, but that researchers may not want to include in their analysis.

1.2.3 Data Comparability

The unstructured, complex, and varied nature of qualitative data can make it difficult to analyze an archived qualitative dataset so as to yield a meaningful answer to a new research question. For big social research, data may have different file types, 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 can 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 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, broad consent (that is, consent to data reuse) has begun to be included in consent forms, and Institutional Review Boards (IRBs) can provide guidelines for consent procedures that allow the use of qualitative data beyond its original purpose. On the other hand, big social researchers often consider big social data to be content that is simply found online, and therefore may not consider it necessary to obtain informed consent from the users who generate big social data. Big social researchers may also consider it sufficient that users have agreed to their social media platforms' terms of service; these terms generally include consent for different types of 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

Both qualitative researchers who share or reuse 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.

1.2.6 Intellectual Property and Data Ownership

Intellectual property 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, the data may be owned by institutions, or intellectual property rights may be held by research participants. In either case, consent from intellectual property rights holders is necessary to redistribute the data for reuse. For big social data, the intellectual property 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 Address Issues in Qualitative Data Reuse and Big Social Research

The rapidly evolving data landscape presents interesting possibilities for social and behavioral research. And as more researchers share data and conduct big social research, there is an increased need for assistance in responsible big social 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 reuse and big social research 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 epistemological, ethical, and legal issues that arise when using data originally created through interaction with human subjects.

A growing body of literature suggests that data curation strategies can alleviate some of the epistemological, ethical, and legal issues described above. These practices include data management planning, designing research to facilitate later data sharing, and producing metadata and other documentation to capture contextual information. Data curation

strategies can also help protect participants from harm, through data deidentification, aggregating data, or restricting access to data. Data curation for qualitative data reuse is a more established practice, and literature going back to the 1990s examines how data curation strategies 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 conducting research, encouraging transparency, and protecting research subjects.

1.4 Goal and Structure of the Book

This book suggests that comparing data curation practices for qualitative data reuse and big social research can help researchers responsibly scale up their research practices. By exploring the similarities and differences between the epistemological, ethical, and legal issues in qualitative data reuse and big social research, this book identifies data curation strategies that can encourage responsible use and reuse of qualitative data, both big and small. These strategies 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 use and reuse of these data.

The book is divided into eight chapters, including this introduction. Chapter 2 outlines my general theoretical approach to the research, provides a brief summary of my research methods, and defines common terms that are used throughout the book. Chapters 3 and 4 review existing literature in qualitative data reuse and big social research; through these literature reviews, I identify the six key issues outlined above—context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. Chapter 5 explores the similarities and differences between these key issues in qualitative data reuse and big social research, especially focusing on the data curation implications of these issues. Chapter 6 provides a detailed description of interviews with qualitative researchers, big social researchers, and data curators. Chapter 7 synthesizes, proposes recommendations, and suggests areas of focus for data curators, based on the literature and insights presented in previous chapters. Chapter 8 suggests future work that can continue to enhance responsible practices when scaling up social and behavioral research, and presents concluding thoughts about the role of data curation in facilitating epistemologically sound, ethical, and legal qualitative data reuse and big social research.

References

Cappella JN (2017) Vectors into the future of mass and interpersonal communication research: big data, social media, and computational social science. Hum Commun Res 43:545–558. https://doi.org/10.1111/hcre.12114

Carroll SR, Herczog E, Hudson M, Russell K, Stall S (2021) Operationalizing the CARE and FAIR principles for indigenous data futures. Sci Data 8:108. https://doi.org/10.1038/s41597-021-00892-0

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 Qual Soc Res 6. https://doi.org/10.17169/fqs-6.1.498

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. Qual Quant 1–14. https://doi.org/10.1007/s11135-018-0757-y

Fan W, Gordon MD (2014) The power of social media analytics. Commun ACM 57:74–81. https://doi.org/10.1145/2602574

FNIGC (2010) The first nations principles of OCAP®, a registered trademark of the First Nations Information Governance Centre (FNIGC). First Nations Information Governance Centre, Akwesasne, ON

Glenna L, Hesse A, Hinrichs C, Chiles R, Sachs C (2019) Qualitative research ethics in the big data era. Am Behav Sci 63:555–559. https://doi.org/10.1177/0002764219826282

Halford S, Savage M (2017) Speaking sociologically with big data: symphonic social science and the future for big data research. Sociology 51:1132–1148. https://doi.org/10.1177/0038038517698639

Manovich L (2012) Trending: the promises and the challenges of big social data. In: Gold MK (ed) Debates in the digital humanities. University of Minnesota Press, Minneapolis, MN, pp 460–475

Mason J (2002) Qualitative research resources: a discussion paper. Prepared for the ESRC Research Resources Board (unpublished, obtained from author)

Mneimneh Z, Pasek J, Singh L, Best R, Bode L, Bruch E, Budak C, Davis-Kean P, Donato K, Ellison N, gelman andrew, Groshen E, Hemphill L, Hobbs W, Jensen JB, Karypis G, Ladd JM, O'Hara A, Raghunathan T, Resnik P, Ryan R, Soroka S, Traugott M, West B, Wojcik S (2021) Data acquisition, sampling, and data preparation considerations for quantitative social science research using social media data. PsyArXiv

Resnick B (2016) Researchers just released profile data on 70,000 OkCupid users without permission. Vox

Wenger E, McDermott RA, Snyder W (2002) Cultivating communities of practice: a guide to managing knowledge. Harvard Business School Press, Boston, MA

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18

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. Qual Health Res. https://doi.org/10.1177/1049732318759658

Zimmer M (2010) "But the data is already public": on the ethics of research in Facebook. Ethics Inf Technol 12:313–325. https://doi.org/10.1007/s10676-010-9227-5

Theoretical Approach, Methods, and Definitions

2

To build the foundation for the rest of the book, this chapter provides an overview of my general theoretical approach to this research, provides a summary of my research methods, and then defines key terms that I use throughout the book: *qualitative data*, *qualitative data reuse*, *big social data,* and *big social research.*

2.1 Theoretical Approach

Information science explores multidisciplinary issues, with an aim of understanding how people interact with information. Cronin suggests that the past few decades have brought about a "sociological turn" in information science research, built on a foundation of social constructivism (Cronin 2008). Social constructivism is based on Vygotsky's social constructivist theory of cognitive development, which emphasizes that people form knowledge through a combination of cognitive processes and social environmental factors (Talja et al. 2005).

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 (Garfinkel 1967), diffusion of innovations theory (Rogers 2003) and actor-network theory (Latour 1996). The research presented in this book is also part of the social constructivist paradigm; it is built upon the idea that qualitative researchers, big social researchers, and data curators have developed different practices and viewpoints based on their sociocultural environments. This book explores how researchers and curators have constructed

S. Mannheimer, *Scaling Up: How Data Curation Can Help Address Key Issues in Qualitative Data Reuse and Big Social Research*, Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-49222-8_2

knowledge around data use and reuse, then synthesizes their insights and approaches to support ethical, legal, and epistemologically sound research and data sharing practices.

To further the goal of understanding the communities investigated in this research (qualitative and big social research communities, and the data curation community), this book also incorporates the idea of communities of practice (Lave and Wenger 1991; Wenger 1998). Communities of Practice Theory helps social science researchers group and analyze scientific communities, with a goal of explaining how groups of people disseminate knowledge. Wenger et al. (2002) define communities of practice as "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." This book examines three distinct communities of practice: qualitative researchers who reuse or share data, big social researchers, and data curators.

Each community of practice has three key characteristics: their *domain*, their *community*, and their *practice* (Wenger et al. 2002). *Domain* describes a set of shared interests and disciplines; c*ommunity* forms when those in the domain 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.

Using qualitative researchers as an example, the *domain* is (1) the interests of qualitative researchers—for example, interest in human behavior, human phenomena, and qualitative research methods, and (2) the disciplines that these researchers come from—for example, anthropology, sociology, or health sciences. The *community* forms when qualitative researchers meet at conferences, cite each other's research, or have community calls. The *practice* may include qualitative content analysis, grounded theory, ideas such as "researcher as instrument," and the shared commitment to in-depth research into human behavior.

Communities of practice theory has been widely used in library and information science, including to study science collaboratories (Bos et al. 2007), to build data management and digital scholarship services in academic libraries (Smith et al. 2020), and as a framework for supporting undergraduate student researchers (Pirmann et al. 2023).

2.2 Methods

The research described in this book uses two methods. First, in Chaps. 3, 4 and 5, I review the existing literature in qualitative data reuse and big social research, synthesize key ideas, and identify six issues in common between qualitative data reuse and big social research, with a focus on data curation. Second, in Chaps. 6 and 7, I use the six issues identified in the earlier chapters to inform semi-structured interviews with three different types of participants, referred to throughout this book as *communities of practice*:

- big social researchers who have conducted research with big social data
- qualitative researchers who have shared or reused qualitative data
- data curators who have worked with one or both types of data.

Through a qualitative content analysis of these interviews, I confirm and build upon the six key issues identified in Chaps. 3, 4 and 5, I additionally suggest three new lenses for considering the practices of these communities: domain differences, strategies for responsible practice, and perspectives on data curation and data sharing. The interview methods are discussed further in Chap. 6. For full details about research methods and sampling, as well as interview guides, transcripts, codebook, analysis, and other documentation, please see the associated dataset:

> Mannheimer (2023) Interviews regarding data curation for qualitative data reuse and big social research. Qualitative Data Repository. https://doi.org/10.5064/F6GWMU4O.

2.3 Definitions

Defining key terms will help readers build a foundation for understanding the research in this book. Therefore, in this section, I provide in-depth definitions of *qualitative data*, *qualitative data reuse*, *big social data*, and *big social research.*

2.3.1 Defining Qualitative Data

In the broadest sense, qualitative data (in contrast to quantitative data) are data that are not numeric (Kitchin 2014). To clarify further, while qualitative data may be analyzed to produce numeric results such as code counts and statistics, the foundational qualitative data themselves are non-numeric (Greener 2011; DuBois et al. 2018).

Bernard et al. (1986) define the construction of qualitative 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;" 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." As these passages suggest, qualitative data are produced by qualitative research, and therefore the term *qualitative data* can be defined by the process that was used to create or collect that data. The National Endowment for the Humanities Office of Digital Humanities (2019) corroborates this

Table 2.1 Examples of qualitative data based on form and access

	Public	Private
Physical objects	Street litter, building architecture, sculpture parks, playground equipment	Curios, mementos, home decor, coin collections, lawn ornaments, DNA samples
Text	Craigslist ads, commemorative plaques, books, congressional data, blogs	To-do lists, short-answer survey responses, personal emails, diary entries
Images	Images from magazines, paintings in galleries and museums, billboards, street art	Drawings in a personal sketchbook, family photos, patient scans
Audio	Podcast episodes, interview recordings, focus group recordings, awards show acceptance speeches, songs	Voice memo recordings, voicemail messages, private conversations
Video	YouTube ads, tv shows, TikTok videos, feature films, documentaries	Smartphone videos, VHS home videos, private event footage (e.g., videos of weddings or retirement parties)

idea, defining data as "materials generated or collected during the course of conducting research."

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). To meet the aims described by Corti, qualitative researchers collect and examine various types of data. Bernard, Wutich, and Ryan (2017) suggest that qualitative data exist in five formats: (1) physical objects, (2) still images, (3), sounds, (4) moving images; and (5) texts. In Table 2.1, I provide examples of public data and private data for each of these categories.

The types of data identified in Table 2.1 are far-reaching and include many types of data that a qualitative researcher could analyze. I include a variety of examples, including both analog and digital data, and both small-scale and large-scale data.

Heaton suggests a classification structure for qualitative data that divides these different types of data into "non-naturalistic" data (i.e., data that are solicited by researchers through interviews, questionnaires, etc.), and "naturalistic" data (i.e., data that are found or collected by researchers with minimal interaction with the research subjects) (Heaton 2004). In Table 2.2, I suggest some examples of non-naturalistic and naturalistic qualitative data.

As with the examples of qualitative data listed in Table 2.1, non-naturalistic and naturalistic data can be either analog or digital in format. For example, fieldnotes could take the form of paper notebooks or word processing documents; diaries could be written using pen and paper, kept using a notetaking 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.

Table 2.2 Examples of Non-naturalistic and Naturalistic qualitative data

	Examples
Non-naturalistic qualitative data	Video recordings of Zoom interviews, focus group transcripts, responses to open-ended survey questions, field notes, audio recordings of research-related conversations
Naturalistic qualitative data	Overheard conversations, public documents, websites, social media, archival materials such as diaries, correspondence, or photos

For purposes of this book, taking into account the kinds of data listed in Tables 2.1 and 2.2, 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.

2.3.2 Defining Qualitative Data Reuse

The term *secondary analysis* emerged in the 1950s do describe a research methodology that uses pre-existing data. Lipset and Bendix (1959) provide a simple definition of this concept: "the study of specific problems through analysis of existing data which were originally collected for another purpose."

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 (1976) 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," and Hakim (1982) 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." In her 2004 definition of qualitative secondary analysis, Heaton (2004) 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]." 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; Scheff 1986; Guba and Lincoln 1989). 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). 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 and Mauthner 2004). 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). 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). 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" (Heaton 2004).

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 book 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" (Thorne 2004)—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, the 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 more expansive term "data reuse." Bishop and Kuula-Luumi (2017) suggest that

"reuse provides an opportunity to study the raw materials of past research projects to gain methodological and substantive insights." van de Sandt et al. (2019) take an even broader 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."

One final note: the various definitions reviewed here 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 (Hinds et al. 1997; Thorne 1998; Heaton 2004; Sherif 2018), Mauthner et al. (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, I consider all data reuse to have 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.

Taking all of these existing definitions and conversations into account, and limiting my definition to the scholarly use of data, this book uses the term *qualitative data reuse*, with the following definition:

> Qualitative data reuse is when researchers use existing qualitative data to refine ideas, gain new insights, and produce new scholarship.

2.3.3 Defining Big Social Data

Big data are often defined in terms of three "Vs": volume, velocity, and variety (Laney 2001; Diebold 2012; Zikopoulos 2012; Kitchin 2014). 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. (Boyd and Crawford 2012)

  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, they have 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; Schroeder 2016; Raguseo 2018), healthcare (e.g., Chawla and Davis 2013; Raghupathi and Raghupathi 2014; Viceconti et al. 2015; Wang et al. 2018), education (e.g., Picciano 2014; Williamson 2017; Nazarenko and Khronusova 2017), and journalism (e.g., Gray et al. 2012; Lewis 2015; Borges-Rey 2016).

The term *big social data* (or sometimes *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. Amer-Yahia et al. (2010) differentiate between *direct* and *indirect* human participation in big data. Big data resulting from direct human participation 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). Big data resulting from indirect human participation 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 and Haider 2015). In Table 2.3, I provide examples of different kinds of big social data, informed by on Amer-Yahia et al. (2010), Olshannikova et al. (2017), Yanai (2012), Ramasamy et al. (2013), and Drakonakis et al. (2019).

In addition to the table above, I also present Table 2.4, below. Table 2.4 uses a similar structure to Table 2.1, in Sect. 2.3.1, so as to demonstrate the relationship between big social data and qualitative data. Contrasting Table 2.1 (Examples of qualitative data based

Table 2.3 Examples of direct human interaction data and indirect human interaction data

	Subcategories	Examples
Direct human interaction data	Data related to individual users	Usernames, passwords, tweets, Instagram photos, TikTok videos, tagged photos, @-mentions
	Data related to user communication and dialogue	Direct messages, comments on a news story, Wikipedia editing data, Slack chats, videoconferences
Indirect human interaction data	Data related to user relationships	Followers, likes, views, network analysis data
	Automatically created metadata	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)

Table 2.4 Examples of big social data based on form and access

	Public	Private	Ambiguous
Text	Online obituaries, twitter posts using hashtags, blogs, news stories	Emails, notes taken on notetaking apps, short responses to survey questions	Comments on other people's Twitter posts, online forum posts
Images	Instagram posts from public figures, Flickr images	Personal photos, digital patient scans, Instagram posts from private profiles	Openly accessible Instagram profile posts from non-public figures
Audio	Podcast ads, songs, online news audio clips	Voice memos, voicemail messages, interview recordings	Digital oral histories
Video	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

on form and access) with Table 2.4 (Examples of big social data based on form and access) highlights two notable differences between qualitative data and big social data. First, Table 2.4 does not include "physical objects," because big social data are by nature digital. Second, while Table 2.1 categorizes qualitative data into "public" and "private," Table 2.4 adds a third category, "ambiguous." As Nissenbaum suggests in her theory of contextual integrity (2009), and as is discussed further in Chap. 4, Sect. 4.3.5, 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 2.4, the column labeled "ambiguous" includes examples such as open Instagram posts from non-public figures that may be accessible publicly, but are designed for a limited, private audience.

Social media is a common source for big social data. Here, I use the term *social media* to describe emerging digital technologies associated with Web 2.0 (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 (2007) 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." 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 they allow users to interact within communities and to create and share digital content in a networked environment (Ip and Wagner 2008; Lüders 2008; Kim et al. 2010; Wilson et al. 2011; Bechmann and Lomborg 2012). Bechmann and Lomborg outline

three characteristics that are commonly emphasized when considering social media as a social phenomenon:

1. Social media platforms facilitate 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 and Lomborg 2012)

A fourth consideration is that social media platforms are often controlled by private, for-profit companies (Driscoll and 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) and TikTok, online forums like Reddit (owned by Conde Nast) and Quora, virtual worlds like Facebook's Metaverse, or the communities that form among videogame users—these platforms all act as intermediaries between the human communities that are formed online (Oboler et al. 2012; Fuchs 2017). 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 these data is often controlled by private companies.

2.3.4 Defining Big Social Research

To define big social research, I will begin by outlining two key types of internet-mediated research: *obtrusive* and *unobtrusive*, as defined by Hewson et al. (2016). In Table 2.5, below, I give examples of obtrusive and unobtrusive internet mediated research.

Table 2.5 Examples of obtrusive and unobtrusive internet-mediated research

	Examples
Obtrusive research	• Online experiments in which participants are aware of their participation, such as when social science researchers recruit participants online and use web-based experiment strategies • Surveys and questionnaires that are distributed via email or online links • Interviews and focus groups that are conducted online
Unobtrusive research	• Experiments in which participants are not aware of their participation, such as A/B testing • Digital observation, such as analysis of interactions in online forums and social media sites • Digital document analysis, such as analysis of blogs, email archives, or Flickr photos

These types of internet-mediated research are reminiscent of the two types of qualitative data outlined in Table 2.2, in Sect. 2.3.1.: 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., field notes, 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.

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). 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; Mason et al. 2014; Berkout et al. 2019).

Taking into account the literature and conversations reviewed above, I define big social research as follows:

> Big social research is when researchers use large-scale data from social media or other online social spaces to gain insights and produce scholarship.

2.4 Chapter Summary

The theoretical approach, definitions, and methods presented here provide a foundation for the rest of the book. The definitions of qualitative data reuse and big social research especially begin to demonstrate the shared characteristics and unique qualities of these two types of research. The next two chapters review existing literature to further explore these similarities and differences, identifying key issues that are shared between qualitative data reuse (Chap. 3) and big social research (Chap. 4). The rest of the book continues to compare and contrast qualitative data reuse and big social research, aiming to inform data curation strategies to support epistemologically sound, ethical, and legal data sharing and use.

References

Amer-Yahia S, Doan A, Kleinberg J, Koudas N, Franklin M (2010) Crowds, clouds, and algorithms: exploring the human side of "big data" applications. In: Proceedings of the 2010 ACM SIGMOD international conference on management of data. ACM, Indianapolis Indiana USA, pp 1259–1260

Bankes S, Lempert R, Popper S (2002) Making computational social science effective: epistemology, methodology, and technology. Soc Sci Comput Rev 20:377–388. https://doi.org/10.1177/089443902237317

Bechmann A, Lomborg S (2012) Mapping actor roles in social media: different perspectives on value creation in theories of user participation. New Media Soc 15:765–781. https://doi.org/10.1177/1461444812462853

Berkout OV, Cathey AJ, Kellum KK (2019) Scaling-up assessment from a contextual behavioral science perspective: potential uses of technology for analysis of unstructured text data. J Contextual Behav Sci 12:216–224. https://doi.org/10.1016/j.jcbs.2018.06.007

Bernard HR, Pelto PJ, Werner O, Boster J, Romney AK, Johnson A, Ember CR, Kasakoff A (1986) The construction of primary data in cultural anthropology. Curr Anthropol 27:382–396. https://doi.org/10.1086/203456

Bernard HR, Wutich A, Ryan GW (2017) Analyzing qualitative data: systematic approaches, 2nd edn. Sage Publications, Los Angeles, CA

Bishop L, Kuula-Luumi A (2017) Revisiting qualitative data reuse: a decade on. Sage Open 7. https://doi.org/10.1177/2158244016685136

Borges-Rey E (2016) Unravelling data journalism. J Pract 10:833–843. https://doi.org/10.1080/17512786.2016.1159921

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. J Comput-Mediat Commun 12:652–672. https://doi.org/10.1111/j.1083-6101.2007.00343.x

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

Boyd D, Crawford K (2012) Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inf, Commun Soc 15:662–679. https://doi.org/10.1080/1369118X.2012.678878

Boyd D, Ellison N (2007) Social network sites: definition, history, and scholarship. J Comput Mediat Commun 13:210–230. 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: Fielding NG, Lee RM, Blank G (eds) The sage handbook of online research methods. Sage Publications, London, UK, pp 125–139

Caliandro A (2018) Digital methods for ethnography: analytical concepts for ethnographers exploring social media environments. J Contemp Ethnogr 47:551–578. https://doi.org/10.1177/0891241617702960

Castells M (2000) Materials for an exploratory theory of the network society. Br J Sociol 51:5–24. https://doi.org/10.1111/j.1468-4446.2000.00005.x

Chawla NV, Davis DA (2013) Bringing big data to personalized healthcare: a patient-centered framework. J Gen Intern Med 28:660–665. https://doi.org/10.1007/s11606-013-2455-8

Chen H, Chiang RHL, Storey VC (2012) Business intelligence and analytics: from big data to big impact. MIS Q 36:1165–1188. https://doi.org/10.2307/41703503

Corti L (1999) Text, sound and videotape: the future of qualitative data in the global network. IASSIST Q 23:15. 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: Qual Soc Res 1. https://doi.org/10.17169/fqs-1.3.1019
Cronin B (2008) The sociological turn in information science. J Inf Sci 34:465–475. https://doi.org/10.1177/0165551508088944
Diebold FX (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
Drakonakis K, Ilia P, Ioannidis S, Polakis J (2019) Please forget where I was last summer: the privacy risks of public location (meta) data. In: Proceedings 2019 network and distributed system security symposium. Internet Society, San Diego, CA
Driscoll K, Walker S (2014) Working within a black box: transparency in the collection and production of big Twitter data. Int J Commun 8:20
DuBois JM, Strait M, Walsh H (2018) Is it time to share qualitative research data? Qual Psychol 5:380–393. https://doi.org/10.1037/qup0000076
Fuchs C (2017) Social media: a critical introduction. Sage Publications
Gandomi A, Haider M (2015) Beyond the hype: big data concepts, methods, and analytics. Int J Inf Manage 35:137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007
Garfinkel H (1967) Studies in ethnomethodology. Polity Press, Cambridge, UK
Glass GV (1976) Primary, secondary, and meta-analysis of research. Educ Res 5:3–8. https://doi.org/10.2307/1174772
Gray J, Bounegru L, Chambers L (eds) (2012) The data journalism handbook. European Journalism Centre, Brussels, Belgium
Greener I (2011) Designing social research: a guide for the bewildered. Sage Publications, London, UK
Guba EG (1981) Criteria for assessing the trustworthiness of naturalistic inquiries. Educ Commun Technol 29:75–91. https://www.jstor.org/stable/30219811
Guba EG, Lincoln YS (1989) Fourth generation evaluation. Sage Publications, Thousand Oaks, CA
Hakim C (1982) Secondary analysis in social research : a guide to data sources and methods with examples. Allen and Unwin, London, UK
Hammersley M (1997) Qualitative data archiving: some reflections on its prospects and problems. Sociology 31:131–142. https://doi.org/10.1177/0038038597031001010
Heaton J (2004) Reworking qualitative data. Sage Publications, London, UK
Heaton J (1998) Secondary analysis of qualitative data. Social Research Update 6
Hewson C, Vogel C, Laurent D (2016) Internet-mediated research: state of the art. In: Internet research methods. Sage Publications, London, UK
Hinds PS, Vogel RJ, Clarke-Steffen L (1997) The possibilities and pitfalls of doing a secondary analysis of a qualitative data set. Qual Health Res 7:408–424. https://doi.org/10.1177/104973239700700306
Ip RKF, Wagner C (2008) Weblogging: a study of social computing and its impact on organizations. Decis Support Syst 45:242–250. https://doi.org/10.1016/j.dss.2007.02.004
Irwin S (2013) Qualitative secondary data analysis: ethics, epistemology and context. Prog Dev Stud 13:295–306. https://doi.org/10.1177/1464993413490479
Ishikawa H (2015) Social big data mining. CRC Press, Boca Raton, FL
Kim W, Jeong O-R, Lee S-W (2010) On social web sites. Inf Syst 35:215–236. https://doi.org/10.1016/j.is.2009.08.003
Kitchin R (2014) The data revolution: big data, open data, data infrastructures & their consequences. Sage Publications, Los Angeles, CA
Laney D (2001) 3D data management: controlling data volume, velocity and variety. Meta Group

Latour B (1996) On actor-network theory: a few clarifications. Soziale Welt 47:369–381
Lave J, Wenger E (1991) Situated learning legitimate peripheral participation. Cambridge University Press, Cambridge, UK
Lewis SC (2015) Journalism in an era of big data. Digit J 3:321–330. https://doi.org/10.1080/21670811.2014.976399
Liebowitz J (ed) (2013) Big data and business analytics. Auerbach Publications, New York, NY
Lipset SM, Bendix R (1959) Social mobility in industrial society. University of California Press, Berkeley, CA
Lüders M (2008) Conceptualizing personal media. New Media Soc 10:683–702. https://doi.org/10.1177/1461444808094352
Mannheimer S (2023) Interviews regarding data curation for qualitative data reuse and big social research. Qual Data Repos. https://doi.org/10.5064/F6GWMU4O
Mason W, Vaughan JW, Wallach H (2014) Computational social science and social computing. Mach Learn 95:257–260. https://doi.org/10.1007/s10994-013-5426-8
Mauthner NS, Parry O, Backett-Milburn K (1998) The data are out there, or are they? Implications for archiving and revisiting qualitative data. Sociology 32:733–745. https://doi.org/10.1177/0038038598032004006
Moore N (2007) (Re)using qualitative data? Sociological Research Online 12:1–13. https://doi.org/10.5153/sro.1496
National Endowment for the Humanities (2019) Data management plans for NEH Office of Digital Humanities proposals and awards
Nazarenko MA, Khronusova TV (2017) Big data in modern higher education: benefits and criticism. In: Quality management, transport and information security, information technologies. pp 676–679
Nissenbaum H (2009) Privacy in context: technology, policy, and the integrity of social life. Stanford University Press, Palo Alto, CA
Oboler A, Welsh K, Cruz L (2012) The danger of big data: social media as computational social science. First Monday 17. https://doi.org/10.5210/fm.v17i7.3993
Olshannikova E, Olsson T, Huhtamäki J, Kärkkäinen H (2017) Conceptualizing big social data. Journal of Big Data 4. https://doi.org/10.1186/s40537-017-0063-x
Parry O, Mauthner NS (2004) Whose data are they anyway? Practical, legal and ethical issues in archiving qualitative research data. Sociology 38:139–152. https://doi.org/10.1177/0038038504039366
Paulus T, Warren A, Lester JN (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
Picciano AG (2014) Big data and learning analytics in blended learning environments: benefits and concerns. Int J Interact Multimed Artif Intell 2:35–43. https://doi.org/10.9781/ijimai.2014.275
Pirmann C, Miessler RC, Baugess C, Moore K, Paddick C (2023) Bridging communities of practice: cross-institutional collaboration for undergraduate digital scholars. In: Hensley MK, Fargo H, Davis-Kahl S (eds) Undergraduate research and the academic librarian: case studies and best practices, vol 2. Association of College and Research Libraries. Chicago, IL, pp 83–102
Raghupathi W, Raghupathi V (2014) Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2:3. 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. Int J Inf Manage 38:187–195. https://doi.org/10.1016/j.ijinfomgt.2017.07.008
Ramasamy D, Venkateswaran S, Madhow U (2013) Inferring user interests from tweet times. In: Proceedings of the first ACM conference on online social networks. Association for Computing Machinery, Boston, MA, pp 235–240

Rogers EM (2003) Diffusion of innovations, 5th edn. Free Press, New York, NY
Ruthven I, Buchanan S, Jardine C (2018) Relationships, environment, health and development: the information needs expressed online by young first-time mothers. J Am Soc Inf Sci 69:985–995. https://doi.org/10.1002/asi.24024
Scheff TJ (1986) Toward resolving the controversy over "thick description." Curr Anthropol 27:408–409. https://doi.org/10.1086/203460
Schroeder R (2016) Big data business models: challenges and opportunities. Cogent Soc Sci 2. https://doi.org/10.1080/23311886.2016.1166924
Sherif V (2018) Evaluating preexisting qualitative research data for secondary analysis. Forum Qualitative Sozialforschung/Forum: Qual Soc Res 19. https://doi.org/10.17169/fqs-19.2.2821
Smith PL, Felima C, Durant F, Van Kleeck D, Huet H, Taylor LN (2020) Building socio-technical systems to support data management and digital scholarship in the social sciences. In: Crowder JW, Fortun M, Besara R, Poirier L (eds) Anthropological data in the digital age: new possibilities—new challenges. Springer International Publishing, Cham, Switzerland, pp 31–57
Stenbacka C (2001) Qualitative research requires quality concepts of its own. Manag Decis 39:551–556. https://doi.org/10.1108/EUM0000000005801
Talja S, Tuominen K, Savolainen R (2005) "Isms" in information science: constructivism, collectivism and constructionism. J Doc 61:79–101. https://doi.org/10.1108/00220410510578023
Thorne S (1998) Ethical and representational issues in qualitative secondary analysis. Qual Health Res 8:547–555. https://doi.org/10.1177/104973239800800408
Thorne S (2004) Secondary analysis of qualitative data. The Sage encyclopedia of social science research methods
Tsai AC, Kohrt BA, Matthews LT, Betancourt TS, Lee JK, Papachristos AV, Weiser SD, Dworkin SL (2016) Promises and pitfalls of data sharing in qualitative research. Soc Sci Med 169:191–198. https://doi.org/10.1016/j.socscimed.2016.08.004
van de Sandt S, Dallmeier-Tiessen S, Lavasa A, Petras V (2019) The definition of reuse. Data Sci J 18:22. https://doi.org/10.5334/dsj-2019-022
Viceconti M, Hunter P, Hose R (2015) Big data, big knowledge: big data for personalized healthcare. IEEE J Biomed Health Inf 19:1209–1215. https://doi.org/10.1109/JBHI.2015.2406883
Wang Y, Kung L, Byrd TA (2018) Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technol Forecast Soc Chang 126:3–13. https://doi.org/10.1016/j.techfore.2015.12.019
Wenger E (1998) Communities of practice: learning, meaning, and identity. Cambridge University Press, Cambridge, UK
Wenger E, McDermott RA, Snyder W (2002) Cultivating communities of practice: a guide to managing knowledge. Harvard Business School Press, Boston, MA
Williamson B (2017) Big data in education: the digital future of learning, policy and practice. Sage Publications, London, UK
Wilson DW, Lin X, Longstreet P, Sarker S (2011) Web 2.0: a definition, literature review, and directions for future research
Yanai K (2012) World seer: a realtime geo-tweet photo mapping system. In: Proceedings of the 2nd ACM international conference on multimedia retrieval. Association for Computing Machinery, Hong Kong, China, pp 1–2
Zikopoulos P (2012) Understanding big data: analytics for enterprise class Hadoop and streaming data. McGraw-Hill, New York, NY

Qualitative Data Reuse in Practice

3

3.1 History of Qualitative Data Reuse

The practice of data reuse goes back to the first part of the twentieth 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). 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; Kiecolt and Nathan 1985; Smith 2008), and literature examining the epistemological, ethical, and legal implications of reusing existing quantitative data in the social sciences (e.g., de Lusignan et al. 2007; Goodwin 2012; Duke and Porter 2013; 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" (Glaser 1962). However, despite this early mention, *qualitative* data reuse did not become a common practice until the 1990s (e.g., Thorne 1994; Hammersley 1997; Hinds et al. 1997; Szabo and Strang 1997; Heaton 1998; Mauthner et al. 1998; Corti 1999; Thompson 2000).

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), but a growing faction of researchers, funding agencies, and

S. Mannheimer, *Scaling Up: How Data Curation Can Help Address Key Issues in Qualitative Data Reuse and Big Social Research*, Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-49222-8_3

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 and Austin 1998; Heaton 2004). The National Institutes of Health began to require data sharing plans in its grant proposals in 2003, and updated guidelines went into effect in 2023 (National Institutes of Health 2023); the National Science Foundation introduced a data management plan requirement to support data sharing and reuse in 2011 (National Science Foundation 2011); the White House Office of Science and Technology Policy released a memo calling for a national commitment to data sharing in 2013 (Holdren 2013), and then updated their guidance in 2022 to promote immediate, free access to data from federally funded projects (Nelson 2022). Some private funders such as Wellcome (2017) and Gates Foundation (2015) require data sharing plans. And 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), *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 aiming to curate and make available qualitative data on a national scale. Qualidata was launched in October 1994 (Corti and 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 and 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). Social science-focused data archives such as ICSPR (ICPSR 2022) and Odum Institute Data Archive (Odum Institute for Research in Social Science 2022) also house qualitative data.

3.2 Benefits of Qualitative Data Reuse

Qualitative data reuse has increased in the twenty-first 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).

The scientific benefits of qualitative data sharing include:

- Building new knowledge, new hypotheses, new methodologies, comparative research, and critiquing or strengthening existing theories. 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. 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 and Brewis 2014).
- Providing data for teaching purposes. 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. For example, Jones and Alexander (Jones et al. 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—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, scientific journals, and academic societies, as described above.

Economic benefits include:

- Avoiding duplication of effort and allowing the conservation of time and resources, therefore supporting a higher return on investment. 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 researchers who used these data archives saw increases in research, teaching and

studying efficiency, and these research gains outweighed the costs of establishing and maintaining the data archives (Beagrie and Houghton 2014).

3.3 Issues in Qualitative Data Reuse

In addition to the potential benefits discussed above, qualitative data reuse raises epistemological, ethical, and legal issues. The epistemological issues of context, data quality and trustworthiness, and data comparability are concerns about the scholarly legitimacy and usefulness of the data—how well can future researchers truly understand the data, and can we ensure that research that reuses qualitative data will be credible and conclusive.

The ethical and legal issues of informed consent, privacy and confidentiality, and intellectual property and data ownership are concerns about the rights of research subjects—ensuring that research participants are informed and protected. Researchers are guided by laws, regulations, and ethical frameworks designed specifically for research. These guidelines are built upon the values of academic disciplines 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 and von dem Bussche 2017). Professional working groups such as Force11/COPE Research Data Publishing Ethics working group (Puebla and Lowenberg 2021), and organizations such as the International Data Spaces Association (IDS Association 2022) also point toward an emerging infrastructure to support ethical and legal data practices in qualitative data reuse.

I discuss six key epistemological, ethical, and legal issues below: context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership.

3.3.1 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. Insights are created through not only reviewing the data, but also through a deep knowledge of the research context and research subjects. That is, in qualitative research,

"meaning is made rather than found" (Mauthner et al. 1998). This meaning is made through the data collection process itself—which can be deeply affected by researchers' own cultural experiences, biases, and decision-making processes. Meaning 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). As Hinds et al. (1992) 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." 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. As Broom et al. (2009) suggest, "the idea that data can be neutralized and deposited into an archive, ready to be 'picked up' by others, sits uncomfortably for many." Dale et al. (1988) voice this discomfort, writing, "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." Pasquetto et al. (2019) write that "removing data from their original context necessarily involves information loss" 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. 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 or collaborate with the researchers who originally collected the data (Hinds et al. 1997; Szabo and Strang 1997; Heaton 2008). However, this strategy is impractical for long-term use of data beyond the lifetime of the original researchers, and furthermore, the original researchers themselves may not remember the full context. Mauthner and Parry discuss in several articles the difficulty of maintaining the context of data, even when attempting to reuse data that they themselves had previously collected (Mauthner et al. 1998; Parry and Mauthner 2004; Mauthner and Parry 2009). As Thorne (1994) describes, researchers may "make mental notes" about participants, settings, and other details that may never be documented in field notes or memos, and may be forgotten later.

Hinds et al. (1997) 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. 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 (Corti 1999, 2000; Fielding 2004; van den Berg 2005; Goodwin and O'Connor 2006; Elman and Kapiszewski 2014; Bernard et al. 2017). Metadata and documentation are discussed in more detail in Sect. 3.4.1.

3.3.2 Data Quality and Trustworthiness

Any reuse of qualitative data relies on the data's quality and trustworthiness, 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 and Appelbaum 1991; Yoon 2017). Sherif (2018) advises that "the original data must allow the researcher conducting secondary analysis to understand examined processes, relationships, and subjective meanings." Hinds et al. (1997) suggest reviewing three randomly selected interviews to determine whether the larger dataset can be used to achieve the research goals of any contemplated new study. Stenbacka (2001) suggests looking at four different dimensions when evaluating a dataset for reuse: "validity, reliability, generalizability and carefulness." I further examine these dimensions of data quality below.

Validity may be affected by errors made during the research process—by research subjects, by reporters or recorders of field data, by researchers, or by data coders. Simple mistakes or inaccuracies can occur throughout the process. And systematic errors can be introduced into datasets as a result of bias related to personal identity, political ideology, general personality, or assumptions. Bernard et al. (1986) suggest that "researchers using archival material need actively to consider potential biases and then, whenever possible, test for them." Reliability can be measured by 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). Reliability can also be determined by evaluating the completeness and accuracy of the dataset. Generalizability can be measured partly by examining the breadth and depth of the dataset, to determine whether the data are appropriate for reuse. The idea of generalizability also overlaps with data comparability, which I examine further in Sect. 3.3.3. Carefulness can be demonstrated through thoughtful and thorough documentation. 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 (Giarlo 2013; Frank et al. 2017; Yoon 2017; ICPSR 2019; Yoon and Lee 2019). Data repositories and academic libraries also support trust through certifications such as the CoreTrustSeal Trustworthy Data Repositories Requirements (CoreTrustSeal 2023) and the TRUST principles for digital repositories (Lin et al. 2020). I further discuss metadata and data archiving in Sect. 3.4.

3.3.3 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. Because qualitative data tends to be relatively unstructured, complex, and varied (Heaton 2004), it

can be difficult to fit a primary dataset into a secondary research question. When attempting to compare and combine qualitative datasets, the literature suggests that researchers use three strategies: (1) identify the extent of missing data; (2) identify how well the research questions converge in the primary research and secondary research; and (3) assess the methods used to produce the primary data (Thorne 1994; Hinds et al. 1997; Heaton 2004).

Another challenge for data comparability is that qualitative researchers often use proprietary qualitative data analysis software such as NVivo and Atlas.ti. These proprietary software programs may not be interoperable and could cause challenges for data reuse. Some research has begun to support standardized formats and interoperability (Corti and Gregory 2011; 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 standardized metadata, and by advocating for open source software and interoperable formats (Karcher et al. 2021). Data curation strategies are further discussed in Sect. 3.4.

3.3.4 Informed Consent

Qualitative researchers have long debated whether participants' consent can ever be truly informed, due to the developmental, reflexive nature of research (Parry and Mauthner 2004). 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 (Wiles et al. 2007; Alexander et al. 2020).

Consent for 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 (Master and Resnik 2013; Joly et al. 2015). As Thorne (1994) writes, "there may be especially sensitive instances in which the implied consent of original subjects cannot be presumed." However, Heaton (1998) suggests that reconsent may often prove too difficult: "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." In a later paper, Heaton 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."

A common strategy to support informed consent for data reuse is to include a clause in the consent form detailing any potential future data sharing, also referred to as *broad*

consent (U.S. Department of Health and Human Services 2017). As Hinds et al. (1997) write, "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." Tiered consent (also called flexible consent, line-item consent, or multilayered consent) can be useful for research in which participants consent to data reuse. The tiered consent model provides participants with a wider variety of options for data sharing—for example, opting out of data sharing completely, consenting to restricted data sharing only, or allowing participants the opportunity to review the data prior to sharing (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 et al. (2022) found that many participants who agreed to "data sharing" misunderstood the term to mean dissemination of research results, even though the consent form contained a detailed description of how the research data would be shared.

The General Data Protection Guidelines (GDPR) in the European Union regulate and define the obligation to communicate clearly about data sharing. 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 and von dem Bussche 2017). 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 (U.S. Department of Health and Human Services 2017). While secondary data use is still viewed as exempt from ethics 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 template language that researchers can use to obtain broad consent and thus support data reuse (Lavori et al. 1999; Siminoff 2003; Elman et al. 2018; Cornell Research Services 2022), and in May 2022, NIH released guidance on consent language for data reuse, indicating that such language may increasingly be standardized (NIH 2022).

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

[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). 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.

researchers and research subjects. There is concern that broad consent exposes respondents to uncertain future risks 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 and Parry 2013).

3.3.5 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 real names or using pseudonyms, by removing any potentially identifying specifics about their lives and experiences, or amalgamating or aggregating data (Clark 2006; Garfinkel 2015). However, some qualitative researchers describe challenges that may arise during the deidentification process. I review these challenges below.

A commonly-cited issue is that that "removal of key identifying characteristics of research participants may…compromise the integrity and quality of the data, or even change their meaning" (Parry and Mauthner 2004). On the other hand, if too much contextual information is present in a dataset—exactly the kind of contextual information that is necessary to understand and reuse the data in the first place—the deidentification may be compromised, thus risking deductive disclosure (Tsai et al. 2016; Myers et al. 2020). Other issues that may affect privacy and confidentiality are limited time and financial resources required (Dorr et al. 2006), and potential technical challenges when deidentifying audiovisual data (Marschik et al. 2023). Additionally, deidentification should be conducted especially thoroughly when participants come from vulnerable populations—e.g., 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).

In addition to these limitations, some argue that there are instances in which deidentification may not in fact be desirable (Turnbull 2000; Moore 2012). Moore (2012) 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 anonymization."

Data curators can support deidentification practices by providing resources and services. If deidentification is not possible or desirable, data repositories can also protect privacy and confidentiality by facilitating restrictions to data access and use (Antes et al. 2018). Access controls are discussed further in Sect. 3.4.2.

3.3.6 Intellectual Property and Data Ownership

Intellectual property is a key consideration for qualitative data reuse (Fienberg et al. 1985; Mauthner et al. 1998; Heaton 2004). As the United States statute states, "copyright protection subsists... in original works of authorship fixed in any tangible medium of expression" (17 U.S. Code § 102 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 my coauthors and I write in 2019, "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" (Mannheimer et al. 2019). 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 and Mauthner 2004). To further complicate matters, universities often claim ownership of research data from affiliated researchers (Steneck 2007).

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 (e.g., interviews, demographic data), method of access (i.e., via computer software), and provisions for reference citations in publications and presentations" (Szabo and Strang 1997). 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)

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 noncommercial 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). Such principles provide guidelines for qualitative researchers and communities who contribute to research to engage with "concerns about fairness, trust, and accountability" and enable contributing communities, "as collectives, to have a say in how their data actually gets used" (Carroll et al. 2021).

In a 2021 survey that asked researchers about their data sharing practices, more than half of respondents reported needing help with copyright and licensing (Simons et al. 2021). 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 Sect. 3.4.

3.4 Data Curation to Support Qualitative Data Reuse

The literature published by the qualitative research community and the data curation community discuss a variety of data curation and archiving practices that respond to the issues described above. These practices can be grouped into two main categories: (1) metadata and documentation; (2) data repositories and professional data curation. While the data curation structures and practices described below cannot address every issue, they do demonstrate that qualitative researchers and data curators are developing a set of strategies to facilitate ethical, legal, and with epistemologically sound qualitative data reuse.

3.4.1 Metadata and Documentation

Metadata and contextual information can serve to prevent "serious misinterpretations and biases in analysis" (White 1991), or secondary researchers making "bolder claims than they otherwise might" (Fienberg et al. 1985). Contextual documentation could include field notes, research diaries, correspondence, and methodological information (Corti and Thompson 1998; Fink 2000; Karcher et al. 2021). 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). 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). 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" (Hinds et al. 1997).

Faniel et al. (2019) interviewed and observed researchers to understand data reuse from the reuser's perspective. Faniel et al.'s findings emphasize three types of information to facilitate data reuse: (1) data production information, including information about data collection, specimen and artifact details, 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 and 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. Data repositories can also encourage researchers to augment their data deposits with 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 and Weber 2019). Qualitative Data Repository's data curation 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 and restrictions, publication procedures, and post-publication procedures (Demgenski et al. 2021).

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). 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 2022) 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. Issues that may complicate the application of DDI metadata to qualitative data include "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). 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 2022). 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 (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 a widely-used standard for describing textual documents (TEI Consortium 2022). 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).

3.4.2 Data Repositories and Professional Data Curation

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). However, it is becoming more common for data repositories to be the preferred sharing method. Notably, the data sharing and data management plans required by funders like NSF and NIH generally ask researchers to formally state how the data will be publicly shared, which has driven an increased demand for data curation and data repository services. Beyond funder requirements, 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 (Wilkinson et al. 2016). This includes providing guidance on data documentation, facilitating data licensing, implementing machine-readable metadata, optimizing data records for search and discovery, and ensuring long-term preservation for published datasets (Demgenski et al. 2021). Data repositories can also provide restricted access to datasets that may not be appropriate for public sharing—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).

There are currently more than two thousand data repositories worldwide, according to the Registry of research Data Repositories (Re3data 2023). Some data repositories such as Dryad Digital Repository (Dryad 2023), ICPSR (ICPSR 2022), and Qualitative Data Repository (Center for Qualitative and Multi-Method Inquiry 2023) provide professional curators who work with data depositors to organize data, create metadata, and otherwise support reuse. Academic libraries also provide research data curation services (Tenopir et al. 2014, 2017; Yoon and Schultz 2017). As mentioned in Sect. 3.4.1, the Data Curation Network brings together academic librarians to support curation for institutional data repositories (Johnston et al. 2018). 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 specialized data curation primers on oral history interviews (Pryse et al. 2021), 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 2023). The TRUST Principles are designed to complement the FAIR Principles to support trustworthy practices for archived data (Lin et al. 2020).

3.5 Chapter Summary

The scientific community is increasingly championing research data reuse. Qualitative data sharing and reuse has steadily grown in the late twentieth and early twenty-first 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. Then in Chap. 5, I comparatively review the issues related to qualitative data reuse and big social data research, and I consider how data curation can help mitigate some of the epistemological, ethical, and legal issues that are present with both data types.

References

17 U.S. Code § 102 (1990) Title 17. Copyrights. Chapter 1. Subject matter and scope of copyright

Akers KG, Read KB, Amos L, Federer LM, Logan A, Plutchak TS (2019) Announcing the Journal of the Medical Library Association's data sharing policy. J Med Libr Assoc 107:468–471. https://doi.org/10.5195/jmla.2019.801

Alexander SM, Jones K, Bennett NJ, Budden A, Cox M, Crosas M, Game ET, Geary J, Hardy RD, Johnson JT, Karcher S, Motzer N, Pittman J, Randell H, Silva JA, da Silva PP, Strasser C, Strawhacker C, Stuhl A, Weber N (2020) Qualitative data sharing and synthesis for sustainability science. Nat Sustain 3:81–88. https://doi.org/10.1038/s41893-019-0434-8

Antes AL, Walsh HA, Strait M, Hudson-Vitale CR, DuBois JM (2018) Examining data repository guidelines for qualitative data sharing. J Empir Res Hum Res Ethics 13:61–73. https://doi.org/10.1177/1556264617744121

APA Data Sharing Working Group (2015) Data sharing: principles and considerations for policy development. American Psychological Association

ASA (2018) American Sociological Association code of ethics

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

Bernanke BS (2004) Editorial statement. Am Econ Rev 94:404

Bernard HR, Pelto PJ, Werner O, Boster J, Romney AK, Johnson A, Ember CR, Kasakoff A (1986) The construction of primary data in cultural anthropology. Curr Anthropol 27:382–396. https://doi.org/10.1086/203456

Bernard HR, Wutich A, Ryan GW (2017) Analyzing qualitative data: systematic approaches, 2nd edn. Sage Publications, Los Angeles, CA

Bill & Melinda Gates Foundation (2015) Open access policy. Bill & Melinda Gates Foundation

Bishop L (2012) Using archived qualitative data for teaching: practical and ethical considerations. Int J Soc Res Methodol 15:341–350. https://doi.org/10.1080/13645579.2012.688335

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

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

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

Castillo D, Coates H, Narlock M (2021) Qualitative data curation primer. Data Curation Network

Center for Qualitative and Multi-Method Inquiry (2023) Qualitative Data Repository. https://qdr.syr.edu/

Clark A (2006) Anonymising research data (NCRM working paper series). ESRC National Centre for Research Methods, p 23

CoreTrustSeal (2023) Core trustworthy data repositories requirements. https://web.archive.org/web/20230407041240/https://www.coretrustseal.org/

Cornell Research Services (2022) IRB consent form templates. https://web.archive.org/web/20220612040831/https://researchservices.cornell.edu/forms/irb-consent-form-templates

Corral M (2020) Atlas.ti data curation primer. Data Curation Network

Corti L (1999) Text, sound and videotape: the future of qualitative data in the global network. IASSIST Q 23:15. 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 Qual Sozialforsch/Forum: Qual Soc Res 1. https://doi.org/10.17169/fqs-1.3.1019
Corti L, Thompson P (1996) ESRC Qualitative Data Archival Resource Centre (QUALIDATA). Sociological Research Online
Corti L, Thompson P (1998) Are you sitting on your qualitative data? Qualidata's mission. Int J Soc Res Methodol 1:85–89. https://doi.org/10.1080/13645579.1998.10846865
Corti L, Backhouse G (2005) Acquiring qualitative data for secondary analysis. Forum Qual Sozialforsch/Forum: Qual Soc Res 6. https://doi.org/10.17169/fqs-6.2.459
Corti L, Gregory A (2011) CAQDAS comparability: what about CAQDAS data exchange? Forum Qual Sozialforsch/Forum: Qual Soc Res 12. https://doi.org/10.17169/FQS-12.1.1634
Cox AM, Kennan MA, Lyon L, Pinfield S (2017) Developments in research data management in academic libraries: towards an understanding of research data service maturity. J Am Soc Inf Sci 68:2182–2200. https://doi.org/10.1002/asi.23781
Dale A, Arber S, Procter M (1988) Doing secondary analysis. Allen & Unwin, Crows Nest, Australia
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
DDI Alliance (2022) 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. Int J Med Informatics 76:261–268. 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. J eSci Librariansh 10:1207. https://doi.org/10.7191/jeslib.2021.1207
di Gregorio S (2019) Unlocking the power of qualitative data for future generations. In: QSR International, NVivo blog. http://www.qsrinternational.com/nvivo/nvivo-community/the-nvivo-blog/unlocking-the-power-of-qualitative-data-for-future. Accessed 8 Nov 2019
Dorr DA, Phillips WF, Phansalkar S, Sims SA, Hurdle JF (2006) Assessing the difficulty and time cost of de-identification in clinical narratives. Methods Inf Med 45:246–252
Dryad (2023) Dryad Digital Repository. https://web.archive.org/web/20230131222439/https://datadryad.org/stash
Dryad Digital Repository (2011) Joint data archiving policy. http://wiki.datadryad.org/Joint_Data_Archiving_Policy_(JDAP)
DuBois JM, Strait M, Walsh H (2018) Is it time to share qualitative research data? Qual Psychol 5:380–393. https://doi.org/10.1037/qup0000076
Duke CS, Porter JH (2013) The ethics of data sharing and reuse in biology. Bioscience 63:483–489. https://doi.org/10.1525/bio.2013.63.6.10
Dunn CS, Austin EW (1998) Protecting confidentiality in archival data resources. IASSIST Q 22:16. https://doi.org/10.29173/iq724
Ellard-Gray A, Jeffrey NK, Choubak M, Crann SE (2015) Finding the hidden participant: solutions for recruiting hidden, hard-to-reach, and vulnerable populations. Int J Qual Methods 14:1–10. https://doi.org/10.1177/1609406915621420
Elman C, Kapiszewski D (2014) Data access and research transparency in the qualitative tradition. PS: Polit Sci Polit 47:43–47. https://doi.org/10.1017/S1049096513001777
Elman C, Kapiszewski D, Vinuela L (2010) Qualitative data archiving: rewards and challenges. PS: Polit Sci Polit 43:23–27. https://doi.org/10.1017/S104909651099077X
Elman C, Kapiszewski D, Lupia A (2018) Transparent social inquiry: implications for political science. Annu Rev Polit Sci 21:29–47. https://doi.org/10.1146/annurev-polisci-091515-025429

Ember CR (2007) Using the HRAF collection of ethnography in conjunction with the Standard Cross-Cultural Sample and the Ethnographic Atlas. Cross-Cult Res 41:396–427. https://doi.org/10.1177/1069397107306593

Evers JC (2018) Current issues in qualitative data analysis software (QDAS): a user and developer perspective. Qual Rep 23:61–73. https://doi.org/10.46743/2160-3715/2018.3205

Evers J, Caprioli MU, Nöst S, Wiedemann G (2020) What is the REFI-QDA standard: experimenting with the transfer of analyzed research projects between QDA software. Forum Qual Sozialforsch/Forum: Qual Soc Res 21. https://doi.org/10.17169/FQS-21.2.3439

Faniel IM, Frank RD, Yakel E (2019) Context from the data reuser's point of view. J Doc. https://doi.org/10.1108/JD-08-2018-0133

Fielding N (2004) Getting the most from archived qualitative data: epistemological, practical and professional obstacles. Int J Soc Res Methodol 7:97–104. https://doi.org/10.1080/1364557031000164069 9

Fienberg SE, Martin ME, Straf ML (eds) (1985) Sharing research data. The National Academies Press, Washington, DC

Fink AS (2000) The role of the researcher in the qualitative research process. A potential barrier to archiving qualitative data. Forum Qual Sozialforsch/Forum: Qual Soc Res 1. 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, Akwesasne, ON

Frank RD, Chen Z, Crawford E, Suzuka K, Yakel E (2017) Trust in qualitative data repositories. Proc Assoc Inf Sci Technol 54:102–111. https://doi.org/10.1002/pra2.2017.14505401012

Garfinkel SL (2015) De-identification of personal information. National Institute of Standards and Technology

Giarlo M (2013) Academic libraries as data quality hubs. J Libr Sch Commun 1:eP1059. https://doi.org/10.7710/2162-3309.1059

Glaser BG (1962) Secondary analysis: a strategy for the use of knowledge from research elsewhere. Soc Probl 10:70–74. https://doi.org/10.2307/799409

Glaser BG (1963) Retreading research materials: the use of secondary analysis by the independent researcher. Am Behav Sci 6:11–14. https://doi.org/10.1177/000276426300601003

Goodwin J (2012) Sage secondary data analysis. Sage Publications, London, UK

Goodwin J, O'Connor H (2006) Contextualising the research process: using interviewer notes in the secondary analysis of qualitative data. Qual Rep 2

Hadley H (2020) NVivo data curation primer. Data Curation Network

Hakim C (1982) Secondary analysis in social research: a guide to data sources and methods with examples. Allen & Unwin, London, UK

Hammersley M (1997) Qualitative data archiving: some reflections on its prospects and problems. Sociology 31:131–142. https://doi.org/10.1177/0038038597031001010

Hartter J, Ryan SJ, MacKenzie CA, Parker JN, Strasser CA (2013) Spatially explicit data: stewardship and ethical challenges in science. PLoS Biol 11:e1001634. https://doi.org/10.1371/journal.pbio.1001634

Heaton J (1998) Secondary analysis of qualitative data. Soc Res Updat 6

Heaton J (2004) Reworking qualitative data. Sage Publications, London, UK

Heaton J (2008) Secondary analysis of qualitative data: an overview. Hist Soc Res/Hist Sozialforsch 33. https://www.jstor.org/stable/20762299

Hinds PS, Chaves DE, Cypess SM (1992) Context as a source of meaning and understanding. Qual Health Res 2:61–74. https://doi.org/10.1177/104973239200200105

Hinds PS, Vogel RJ, Clarke-Steffen L (1997) The possibilities and pitfalls of doing a secondary analysis of a qualitative data set. Qual Health Res 7:408–424. https://doi.org/10.1177/104973239700700306

Hirtle PB, Hudson E, Kenyon AT (2009) Copyright and cultural institutions: guidelines for digitization for U.S. libraries, archives, and museums. Cornell University Library

Holdren JP (2013) Increasing access to the results of federally funded scientific research. White House Office of Science and Technology Policy

Hyman HH (1972) Secondary analysis of sample surveys: principles, procedures, and potentialities. Wiley, New York, NY

ICPSR (2012) Guide to social science data preparation and archiving: introduction. http://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/

Johnston LR, Carlson J, Hudson-Vitale C, Imker H, Kozlowski W, Olendorf R, Stewart C, Blake M, Herndon J, McGeary TM, Hull E (2018) Data Curation Network: a cross-institutional staffing model for curating research data. Int J Digit 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:e0129893. https://doi.org/10.1371/journal.pone.0129893

Jones K, Alexander SM, 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, Smith J, Smorul M, Strasser C, Strawhacker C, Stuhl A, Weber N, Winslow D (2018) Qualitative data sharing and re-use for socio-environmental systems research: a synthesis of opportunities, challenges, resources and approaches. https://doi.org/10.13016/M2WH2DG59

Kansa SW, Kansa EC (2018) Data beyond the archive in digital archaeology: an introduction to the special section. Adv Archaeol Pract 6:89–92. https://doi.org/10.1017/aap.2018.7

Karcher S, Weber N (2019) Annotation for transparent inquiry: transparent data and analysis for qualitative research. IASSIST Q 43:1–9. https://doi.org/10.29173/iq959

Karcher S, Kirilova D, Weber N (2016) Beyond the matrix: repository services for qualitative data. IFLA J 42:292–302. https://doi.org/10.1177/0340035216672870

Karcher S, Kirilova D (Dessi), Pagé C, Weber N (2021) How data curation enables epistemically responsible reuse of qualitative data. Qual Rep 26:1996–2010. https://doi.org/10.46743/2160-3715/2021.5012

Kiecolt JK, Nathan LE (1985) Secondary analysis of survey data. Sage Publications, Los Angeles, CA

Lavori PW, Sugarman J, Hays MT, Feussner JR (1999) Improving informed consent in clinical trials: a duty to experiment. Control Clin Trials 20:187–193. 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. Qual Health Res 11:693–705. 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. https://doi.org/10.17169/fqs-1.3.1025

Lin D, Crabtree J, Dillo I, Downs RR, Edmunds R, Giaretta D, De Giusti M, L'Hours H, Hugo W, Jenkyns R, Khodiyar V, Martone ME, Mokrane M, Navale V, Petters J, Sierman B, Sokolova DV, Stockhause M, Westbrook J (2020) The TRUST Principles for digital repositories. Sci Data 7:144. https://doi.org/10.1038/s41597-020-0486-7

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. Am Behav Sci 63:643–664. https://doi.org/10.1177/0002764218784991

Marschik PB, Kulvicius T, Flügge S, Widmann C, Nielsen-Saines K, Schulte-Rüther M, Hüning B, Bölte S, Poustka L, Sigafoos J, Wörgötter F, Einspieler C, Zhang D (2023) Open video data sharing in developmental science and clinical practice. iScience 26:106348. https://doi.org/10.1016/j.isci.2023.106348

Master Z, Resnik DB (2013) Incorporating exclusion clauses into informed consent for biobanking. Camb Q Healthc Ethics 22:203–212. https://doi.org/10.1017/S0963180112000576

Mauthner NS (2012) 'Accounting for our part of the entangled webs we weave': ethical and moral issues in digital data sharing. Ethics in qualitative research. Sage Publications, London, UK, pp 157–175

Mauthner NS, Parry O (2009) Qualitative data preservation and sharing in the social sciences: on whose philosophical terms? Aust J Soc Issues 44:291–307. https://doi.org/10.1002/j.1839-4655.2009.tb00147.x

Mauthner NS, Parry O (2013) Open access digital data sharing: principles, policies and practices. Soc Epistemol 27:47–67. https://doi.org/10.1080/02691728.2012.760663

Mauthner NS, Parry O, Backett-Milburn K (1998) The data are out there, or are they? Implications for archiving and revisiting qualitative data. Sociology 32:733–745. https://doi.org/10.1177/0038038598032004006

McCall RB, Appelbaum MI (1991) Some issues of conducting secondary analyses. Dev Psychol 27:911–917. https://doi.org/10.1037/0012-1649.27.6.911

Miles MB, Huberman AM, Saldana J (2020) Qualitative data analysis: a methods sourcebook, 4th edn. Sage Publications, Los Angeles, CA

Moore N (2012) The politics and ethics of naming: questioning anonymisation in (archival) research. Int J Soc Res Methodol 15:331–340. https://doi.org/10.1080/13645579.2012.688330

Murdock GP (1961) Outline of cultural materials. Human Relations Area Files, New Haven, CT

Myers CA, Long SE, Polasek FO (2020) Protecting participant privacy while maintaining content and context: challenges in qualitative data de-identification and sharing. Proc Assoc Inf Sci Technol 57:e415. https://doi.org/10.1002/pra2.415

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 Institutes of Health (2023) Final NIH policy for data management and sharing. https://web.archive.org/web/20230308170610/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

Nelson A (2022) Ensuring free, immediate, and equitable access to federally funded research. White House Office of Science and Technology Policy

NIH (2022) Informed consent for secondary research with data and biospecimens: points to consider and sample language for future use and/or sharing

Odum Institute for Research in Social Science (2022) Odum Institute data archive. https://web.archive.org/web/20221212004213/https://odum.unc.edu/archive/

Office for Human Research Protections (2018) Revised Common Rule Q&As. HHS.gov

Parry O, Mauthner NS (2004) Whose data are they anyway? Practical, legal and ethical issues in archiving qualitative research data. Sociology 38:139–152. https://doi.org/10.1177/0038038504039366

Pasquetto IV, Borgman CL, Wofford MF (2019) Uses and reuses of scientific data: the data creators' advantage. Harv Data Sci Rev 1. https://doi.org/10.1162/99608f92.fc14bf2d

PLOS (2014) Data availability. https://web.archive.org/web/20200523181932/https://journals.plos.org/plosone/s/data-availability

Pryse JA, Harp M, Mannheimer S, Marsolek W, Cowles W (2021) Oral history interviews data curation primer. Data Curation Network

Puebla I, Lowenberg D (2021) Joint FORCE11 & COPE Research Data Publishing Ethics working group recommendations. Zenodo. https://doi.org/10.5281/ZENODO.5391293

Re3data (2023) Registry of Research Data Repositories. https://doi.org/10.17616/R3D

Rothstein MA (2010) Is deidentification sufficient to protect health privacy in research? Am J Bioeth 10:3–11. https://doi.org/10.1080/15265161.2010.494215

Sherif V (2018) Evaluating preexisting qualitative research data for secondary analysis. Forum Qual Sozialforsch/Forum: Qual Soc Res 19. https://doi.org/10.17169/fqs-19.2.2821

Siminoff LA (2003) Toward improving the informed consent process in research with humans. IRB: Ethics Hum Res 25:S1–S3. https://doi.org/10.2307/3564115

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. Digit Sci Rep. https://doi.org/10.6084/m9.figshare.17061347.v1

Smith E (2008) Using secondary data in educational and social research. McGraw-Hill Education, Berkshire, UK

State of California (2020) Title 11. Law - division 1. Attorney General - chapter 20. California Consumer Privacy Act Regulations

State of Vermont (2018) H.764 (Act 171)

Stenbacka C (2001) Qualitative research requires quality concepts of its own. Manag Decis 39:551–556. https://doi.org/10.1108/EUM0000000005801

Steneck NH (2007) Chapter 6. Data management practices

Sweeney L, Crosas M, Bar-Sinai M (2015) Sharing sensitive data with confidence: the datatags system. Technol Sci 2015101601. https://web.archive.org/web/20220122022200/https://techscience.org/a/2015101601/

Szabo V, Strang VR (1997) Secondary analysis of qualitative data. Adv Nurs Sci 20:66. https://doi.org/10.1097/00012272-199712000-00008

Taichman DB, Sahni P, Pinborg A, Peiperl L, Laine C, James A, Hong S-T, Haileamlak A, Gollogly L, Godlee F, Frizelle FA, Florenzano F, Drazen JM, Bauchner H, Baethge C, Backus J (2017) Data sharing statements for clinical trials. BMJ 357. https://doi.org/10.1136/bmj.j2372

TEI Consortium (2022) TEI: Text Encoding Initiative. https://web.archive.org/web/20220404150605/https://tei-c.org/

Tenopir C, Sandusky RJ, Allard S, Birch B (2014) Research data management services in academic research libraries and perceptions of librarians. Libr Inf Sci Res 36:84–90. 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 RJ, Allard S (2017) Research data services in European academic research libraries. LIBER Q 27:23–44. https://doi.org/10.18352/lq.10180

Thompson P (2000) Re-using qualitative research data: a personal account. Forum Qual Sozialforsch/Forum: Qual Soc Res 1. https://doi.org/10.17169/fqs-1.3.1044

Thorne S (1994) Secondary analysis in qualitative research: issues and implications. In: Morse JM (ed) Critical issues in qualitative research methods. Sage Publications, London, UK, pp 263–279

Tiffin N (2018) Tiered informed consent: respecting autonomy, agency and individuality in Africa. BMJ Glob Health 3:e001249. https://doi.org/10.1136/bmjgh-2018-001249

Tsai AC, Kohrt BA, Matthews LT, Betancourt TS, Lee JK, Papachristos AV, Weiser SD, Dworkin SL (2016) Promises and pitfalls of data sharing in qualitative research. Soc Sci Med 169:191–198. https://doi.org/10.1016/j.socscimed.2016.08.004

Turnbull A (2000) Collaboration and censorship in the oral history interview. Int J Soc Res Methodol 3:15–34. https://doi.org/10.1080/136455700294905

UK Data Archive (2022) Metadata standards: QuDEx. https://web.archive.org/web/20220302004944/https://www.data-archive.ac.uk/managing-data/standards-and-procedures/metadata-standards/

U.S. Department of Health and Human Services (1991) Federal policy for the protection of human subjects ("Common rule"). HHS.gov

U.S. Department of Health and Human Services (2017) Attachment C - recommendations for broad consent guidance

van den Berg H (2005) Reanalyzing qualitative interviews from different angles: the risk of decontextualization and other problems of sharing qualitative data. Hist Soc Res/Hist Sozialforsch 6: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. Qual Health Res 32:182–191. https://doi.org/10.1177/10497323211054058

Voigt P, von dem Bussche A (2017) The EU General Data Protection Regulation (GDPR). Springer International Publishing, Cham, Switzerland

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

White DR (1991) Sharing anthropological data with peers and Third World hosts. In: Sieber JE (ed) Sharing social science data: advantages and challenges. Sage Publications, pp 42–60

Wiles R, Crow G, Charles V, Heath S (2007) Informed consent and the research process: following rules or striking balances? Sociol Res Online 12:1–12. https://doi.org/10.5153/sro.1208

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18

World Medical Association (2013) Declaration of Helsinki

Wutich A, Brewis A (2014) Food, water, and scarcity: toward a broader anthropology of resource insecurity. Curr Anthropol 55:444–468. https://doi.org/10.1086/677311

Yoon A (2017) Data reusers' trust development. J Am Soc Inf Sci 68:946–956. https://doi.org/10.1002/asi.23730

Yoon A, Schultz T (2017) Research data management services in academic libraries in the U.S.: a content analysis of libraries' websites. Coll Res Libr 78. https://doi.org/10.5860/crl.78.7.920

Yoon A, Lee YY (2019) Factors of trust in data reuse. Online Inf Rev. https://doi.org/10.1108/OIR-01-2019-0014

Big Social Research in Practice

4

4.1 History of Big Social Research

Big social research can be traced back to social network analyses in the early part of the twentieth century (Moreno 1934; Simmel 1955; Halavais 2015). As archived social science data became more common, these data were used to support larger-scale longitudinal studies (Holland et al. 2006; Neale and 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 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 (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

S. Mannheimer, *Scaling Up: How Data Curation Can Help Address Key Issues in Qualitative Data Reuse and Big Social Research*, Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-49222-8_4

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). 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), and machine learning and social network analysis have been used to understand political homophily on Twitter (Colleoni et al. 2014). Other big social researchers have mined hashtags to investigate how Twitter is used as a community organizing tool (Segerberg and Bennett 2011). A systematic review of big social research in environmental science highlighted both major benefits—"unprecedented opportunities to extend the scope, scale and depth of research" into human interaction with the environment, and the risks—"a range of issues involving potential biases, big data management, and rapidly evolving frameworks with which [environmental researchers] are generally not familiar" (Ghermandi and Sinclair 2019). Big social data have 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).

4.2 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). 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 (Lazer et al. 2009; Oboler et al. 2012; Fan and Gordon 2014; Cappella 2017). 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" (Baram-Tsabari et al. 2017). 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). Building off 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 (Moorhead et al. 2013; Taylor and Pagliari 2018). 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 (Hökby et al. 2016; Baram-Tsabari et al. 2017). Additionally, big social data are cost-effective (Chang et al. 2014). 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). Lastly, some argue that big social research is less likely to reflect certain types of 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 (McKee 2013; Taylor and Pagliari 2018). According to Baram-Tsabari et al (2017), "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."

All of these benefits support the increasing use of big social data to investigate human behavior. However, big social data also present several issues and challenges. Boyd and Crawford's inclusion of "mythology" in their definition of big data (see Chap. 2, Sect. 2.3.3) addresses the widespread embrace of big data as a knowledge source. In fact, Boyd and Crawford (2012) 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.'" Kitchin (2014) 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."

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." 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). Section 4.3. discusses these and additional concerns in more detail.

4.3 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." In other words, big data are 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." Proferes' (2017) response to Puschmann further outlines the idea that data cannot "speak for themselves." Citing Barad (2003), Proferes argues that "techno-scientific discursive practices involving language, measurement, and materiality *produce* phenomena, creating an artificial separation between researcher and the knowable." Manovich (2012) 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.

As Boyd and Crawford (2012) write, the advent of big data represents "a profound change at the levels of epistemology and ethics." From the literature, I identify six key epistemological, ethical, and legal issues that align with issues in qualitative data reuse that I identified in Chap. 3, Sect. 3.3: context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. Recent workshops with big social researchers conducted by Clark et al. (2019) confirm these issues; Clark et al. also suggest opportunities and challenges of data sharing, which I discuss throughout this book, and in particular in Sect. 4.4.

Big social research may fall outside of the traditional protections 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 governed by ethics regulatory bodies such as institutional review boards (IRBs). The 2018 revision of the Common Rule declined to regulate big social data, and IRBs in the United States have yet to come to a unified conclusion about ethical standards for big social research (Cooky et al. 2018). As Clark et al. (2019) 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." 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 (Wilkinson and Thelwall 2011). However, in the decade since, the human element of big social data has increasingly been recognized (Shilton and Sayles 2016; Metcalf and Crawford 2016; Zimmer 2018; Franzke et al. 2020).

In 2013, the U.S. Department of Health and Human Services 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" (Secretary's Advisory Committee on Human Research Protections 2013). As Buchanan (2017) 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." 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). 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, suggesting that the Office of Human Research Protections (OHRP) could work with IRBs on consent waiver standards for big data research, and on strategies 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. (2018) outline several issues regarding big social research that "may not be adequately covered by existing [ethical] guidelines." 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."

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 and von dem Bussche 2017). While GDPR is a step forward, the ramifications for big social research are still not fully clear (Vestoso 2018; Greene et al. 2019).

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 detail key issues below.

4.3.1 Context

As Halavais (2015) writes, "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. The text, images, audio, or video that are collected as big social data are taken from a larger context of personal and public life (Törnberg and Törnberg 2018), and 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." 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 and 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 have 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 and Nolin 2017). Data combining and data comparability are discussed further in Sect. 4.3.3.

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 online, "multiple audiences [are flattened] into one" (Marwick and Boyd 2011). Users may, in effect, post into a contextual void. Marwick and Boyd suggest that people who post on social media attempt to represent the various facets of their lives and identities to a diverse online community. They therefore may "adopt a variety of tactics, such as using multiple accounts, pseudonyms, and nicknames, and creating 'fakesters' to obscure their real identities" (Marwick and Boyd 2011). This varied self-presentation complicates the idea of authenticity and data quality, as discussed further below.

4.3.2 Data Quality and Trustworthiness

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 et al. (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). Many scholars have also cited Goffman's idea of the presentation of self (1959) 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 and 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. (2015) 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." 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—could introduce errors, bias, and distortion into studies with big social data, and could 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 (Wilson et al. 2012; Zimmer and Proferes 2014; Rains and Brunner 2015; Stoycheff et al. 2017), which could lead to biased research. As Boyd and Crawford (2012) 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." 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." Some researchers have attempted to create more representative datasets by blending big social data with smaller social datasets, as a way to include new perspectives that can help the data be more

meaningful (Croeser and Highfield 2020). Data combining and data comparability are discussed further below.

4.3.3 Data Comparability

Big social researchers may compare and combine data to enhance the representativeness of the data, enhance the context of the data, and achieve stronger results (Croeser and Highfield 2020). 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 (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, it may have different filetypes, different metadata fields, and different metadata standards, all of which make combining data more difficult, especially on a large scale. An additional discussion of data comparability and interoperability as they specifically relate to metadata is included in Sect. 4.4.1.

4.3.4 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 the users' consent is actually *informed* (Obar and Oeldorf-Hirsch 2020). 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). While big social research may carry less risk to participants than other types of research, consent is still an important consideration in any research that involves human participants. As Metcalf (2016) 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." 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). However, these strategies are still largely untested.

Two high-profile cases of research with social media have brought big social 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). This study ultimately prompted 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 Facebook friends. This system 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). To add to the ethical complexity, no Facebook users consented to their data being 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 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 (2013) used pop-up messages to evaluate participants' willingness to share certain types of data on Facebook, 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 and Vahlstrup 2015). However, due to the sheer number of participants in a big social dataset, it is difficult to obtain individual consent, and those who consent may not be fully informed about research risks.

4.3.5 Privacy and Confidentiality

The question of whether data are private or public may blur in online contexts. Manovich (2012) cites Latour (2007), 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." 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 and Bjerrekær 2016), asserting that the data were "already public" and required no special privacy considerations (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" (Beguerisse-Díaz et al. 2017). These are just two examples demonstrating the complex issues and lack of consensus around privacy on social media.

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." 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. (2019) 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." As Marwick and Boyd (2014) 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." Palen and Dourish (2003) 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." 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," and 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 (2014) 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." 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 and Dourish 2003).

Several studies have attempted to understand users' expectations for privacy online. A 2014 Pew Research Center study finds that most respondents wish they could do more to protect their privacy, yet they also believed "it is not possible to be anonymous online" (Madden 2014). Reuter et al. (2019) find that "most users do not think monitoring Twitter for the purpose of clinical trial recruitment constitutes inappropriate surveillance or a violation of privacy." 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." Golder et al. (2019) also conclude that participant responses to social media research vary, depending on "the type of social media platform … the vulnerability of the social media use." Fiesler and Proferes (2018) find that Twitter users have concerns about privacy that align with 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 2023a). However, the privacy settings of social media platforms generally default to open. Users must opt into privacy controls, and implementing such controls may be confusing and difficult (Sleeper et al. 2013).

As in offline research, issues of privacy and confidentiality are especially important when conducting research with participants from vulnerable communities, for whom any potential disclosure poses greater risks (Clark et al. 2019). This is even more true because big social data are used by government entities and advertisers for surveillance. In the 1990s, Sieber (1991) wrote that surveillance "is not a legitimate use of shared data and may be damaging to science." 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). As Oboler, Welsh, and Cruz (2012) write, the ad-driven business model "places the individual's interest in privacy at war with the advertisers' interest in greater customer profiling." 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 Sect. 4.4.2.

4.3.6 Intellectual Property and Data Ownership

Big social research 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 (Doft 2015; Wilkof 2016; Blank 2018; Bosher and Yeşiloğlu 2019).

As noted in Chap. 2, Sect. 2.3.3, 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 and Burgess 2014), and they are a reflection of how much value and revenue are generated through user data. User data help social media companies understand their users, optimize their platforms, and enhance business practices. Social media companies also generate revenue by selling user data to data brokers and advertisers.

Because social media companies view user 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 the case of the OkCupid dataset discussed in Sect. 4.3.5., the dataset was ultimately taken down, partly because of privacy concerns, and also because OkCupid filed a Digital Millennium Copyright Act (DMCA) complaint (McCook 2016). DCMA is commonly used by media companies to flag copyright infringements in online content, such as unlicensed music being played on Instagram live, or unlicensed movies or music being posted on YouTube. The DCMA has been critiqued in the scholarly community as having a chilling effect on innovation by exposing legitimate scholarly, journalistic, and creative activities to potential takedown notices and lawsuits (Lee 2006; Henderson et al. 2007; EFF 2014). However, OkCupid's DCMA complaint appears to be a rare example of the DCMA being evoked in the case of a research dataset. OkCupid's reasoning behind issuing the complaint was not made public.

A more common strategy used in recent years by social media companies is to invoke the Computer Fraud and Abuse Act (CFAA) (the primary federal anti-hacking law) to try to prevent automated web scraping of 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 (938 Federal Reporter 3rd 2019). In that court case, the professional networking platform LinkedIn claimed that the CFAA prohibited the data analytics

company hiQ 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 (141 Supreme Court Reporter 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 were 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" (141 Supreme Court Reporter 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 or via an API, you may only distribute Tweet IDs, Direct Message IDs, and/or User IDs" (Twitter 2023b). 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 to restrict use of data even further (Bruns 2019). For instance, the Twitter Terms of Service for Developers outlines prohibited uses of data and development products as including:

> "Prohibitions on 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 2023c)

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, these restrictions may also limit the topics of study for academic researchers.

4.4 Data Curation to Support Big Social Data 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. The data curation literature outlines a variety of curation and archiving practices that respond to the issues described above. As in Chap. 3, I group these practices into categories: (1) metadata and documentation; (2) data repositories and professional data curation.

4.4.1 Metadata and Documentation

Metadata and documentation can facilitate responsible use and reuse of big social data, and big social data benefits from having embedded descriptive and technical 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). A second kind of metadata can additionally be identified within the text of the tweet: hashtags, @-mentions, and URLs. As Driscoll and Walker (2014) write, "Taken together, these primitive components provide a set of basic descriptive characteristics that might be reported about any collection of tweets." 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

(2016) writes, "Preserving social media means capturing enough content to provide meaning but also finding practical solutions to managing such large, diverse, and interlinked material."

Additionally, the metadata embedded in big social data vary by social media platform. As Acker and Kriesberg (2017) 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." 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 (Schema.org 2020; e.g., DDI Alliance 2022) 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 and Lewis 2014). As I write with coauthor Elizabeth Hull (Mannheimer and Hull 2018), "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." 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 2022), 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.

4.4.2 Data Repositories and Professional Data Curation

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." Driscoll and Walker (2014) 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." 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 Sect. 4.3.6., 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. The Cambridge Analytica scandal also brought an end to what Puschmann (2019) calls the "Wild West of social media research," a period characterized by easy access to big social data, with few rules or regulations. As noted in Sect. 4.3.6, many social media companies changed their data use terms of service and limited API access to their data in response to Cambridge Analytica—a change so swift and disruptive to the status quo that Bruns (2019) deemed it the "APIcolypse." Some social media companies have since formed partnerships with academic institutions that provide structures for academic researchers to gain extended access to data. One example of such a partnership is Social Science One, a partnership between Facebook and researchers at Harvard and Stanford Universities (King and Persily 2020). However, such public–private partnerships still place power in the hands of the social media companies. Public data archiving is a way to bring that power back into the hands of researchers, ensuring open access to big social data for future scholarship.

Weller and Kinder-Kurlanda (2016) 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." 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). 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 less user-friendly access model than the API access provided by social media sites (Ben-David and Huurdeman 2014; Vlassenroot et al. 2019). This leaves 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 (Prom 2017), ICPSR's Social Media Archive (Hemphill et al. 2018), the GESIS

Data Archive (Bishop and Gray 2018), the UK Data Ethics Framework (UK Central Digital and Data Office 2020). Documenting the Now (DocNow) is another project that specifically focuses on "the ethical collection use, and preservation of social media content" (DocNow 2020). DocNow 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. The DocNow 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 (Anderson and Christen 2022), 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 can help protect participant privacy by providing deidentification assistance and disclosure risk reviews. However, the practice of deidentification for big social data is difficult, due to their size, searchability, and potential availability online. 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. (2021) 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."

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 (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). Markham (2012) suggests that deductive disclosure of social media data 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. However, this strategy is more difficult with audiovisual data, and may still not be sufficient to support true deidentification of big social data. Schneble et al. (2018) 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." They also note that "data that are anonymized today might be made re-identifiable tomorrow [through enhanced data technologies]." Metcalf (2016) highlights the unknown risks that may result from 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." These are ongoing challenges that data curators and researchers will continue to face when trying to balance sharing big social data with protecting the people represented in those data.

Another strategy that can support the privacy of the social media users, is 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. Most data repositories provide options for restricted access. Data repositories could also look to existing projects such as the content management system Mukurtu, which was designed specifically to accommodate the different levels of access permissions for digital objects that may be required by Indigenous communities (Christen et al. 2017). 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.

4.5 Summary

The advent of big social data has the potential to reveal large-scale insights about human behavior. However, several key epistemological, ethical, and legal issues arise when conducting research with big social data, as well as when sharing or archiving those data. Data curation practices, including data curation services 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).

In this chapter and Chap. 3, I have presented six issues that are encountered in both 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. In Chap. 5, I review and compare these issues, with an eye toward

using data curation practices as a means to mitigate some of the epistemological, ethical, and legal challenges that are presented by both qualitative data reuse and big social research.

References

141 Supreme Court Reporter (2021) Van Buren v. United States

938 Federal Reporter 3rd (2019) hiQ Labs, Inc v. LinkedIn Corporation

Acker A, Kriesberg A (2017) Tweets may be archived: civic engagement, digital preservation and Obama White House social media data. Proc Assoc Info Sci Tech 54:1–9. https://doi.org/10.1002/pra2.2017.14505401001

Altman I (1977) Privacy regulation: culturally universal or culturally specific? J Soc Issues 33:66–84. https://doi.org/10.1111/j.1540-4560.1977.tb01883.x

Anderson C (2008) The end of theory: the data deluge makes the scientific method obsolete. Wired

Anderson J, Christen K (2022) Local contexts: grounding Indigenous rights. https://web.archive.org/web/20220423194347/https://localcontexts.org/. Accessed 25 Apr 2022

Andrejevic M (2014) Big data, big questions: the big data divide. Inte J Commun 8:17. https://ijoc.org/index.php/ijoc/article/view/2161

Barad K (2003) Posthumanist performativity: toward an understanding of how matter comes to matter. Signs: J Women Culture Soc 28:801–831. https://doi.org/10.1086/345321

Baram-Tsabari A, Segev E, Sharon AJ (2017) What's new? The applications of data mining and big data in the social sciences. The Sage handbook of online research methods. Sage Publications, London, UK, pp 92–106

Barhorst JB, McLean G, Brooks J, Wilson A (2019) Everyday micro-influencers and their impact on corporate brand reputation. In: Proceedings of the 21st ICIG symposium. Durham, England

Bechmann A, Vahlstrup PB (2015) Studying Facebook and Instagram data: the Digital Footprints software. First Monday 20.https://doi.org/10.5210/fm.v20i12.5968

Beguerisse-Díaz M, McLennan AK, Garduño-Hernández G, Barahona M, Ulijaszek SJ (2017) The 'who' and 'what' of #diabetes on Twitter. Digital Health 3.https://doi.org/10.1177/2055207616688841

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

Bishop L, Gray D (2018) Chapter 7: Ethical challenges of publishing and sharing social media research data. In: Woodfield K (ed) The ethics of online research, 1st edn. Emerald Publishing, Bingley, pp 159–188

Blank J (2018) IP law in the age of social media. Northeastern University Graduate Programs

Bond RM, Fariss CJ, Jones JJ, Kramer ADI, Marlow C, Settle JE, Fowler JH (2012) A 61-million-person experiment in social influence and political mobilization. Nature 489:295–298. https://doi.org/10.1038/nature11421

Bosher H, Yeşiloğlu S (2019) An analysis of the fundamental tensions between copyright and social media: the legal implications of sharing images on Instagram. Int Rev Law Comput Technol 33:164–186. 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 Commun Quart 95:471–496. https://doi.org/10.1177/1077699018763307

Boyd d (2013) Bibliography of research on Twitter & microblogging. https://web.archive.org/web/20191123145930/https://www.danah.org/researchBibs/twitter.php. Accessed 23 Nov 2019

Boyd d, Crawford K (2012) Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inf Commun Soc 15:662–679.https://doi.org/10.1080/1369118X.2012.678878

Bright J (2017) 'Big social science': doing big data in the social sciences. In: Fielding NG, Lee RM, Blank G (eds) The Sage handbook of online research methods. Sage Publications, London, UK, pp 125–139

Bruns A (2019) After the 'APIcalypse': social media platforms and their fight against critical scholarly research. Inf Commun Soc 22:1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Bruns A (2013) Faster than the speed of print: reconciling 'big data' social media analysis and academic scholarship. First Monday 18.https://doi.org/10.5210/fm.v18i10.4879

Bruns A, Weller K (2016) Twitter as a first draft of the present: and the challenges of preserving it for the future. In: Proceedings of the 8th ACM conference on web science. Association for computing machinery, hannover, Germany, pp 183–189

Buchanan E (2017) Internet research ethics: twenty years later. In: Zimmer M, Kinder-Kurlanda K (eds) Internet research ethics for the social age: new challenges, cases, and contexts. Peter Lang, New York, NY, pp xxix–xxxiii

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

Cappella JN (2017) Vectors into the future of mass and interpersonal communication research: big data, social media, and computational social science. Hum Commun Res 43:545–558. https://doi.org/10.1111/hcre.12114

Cavazos-Rehg PA, Krauss M, Fisher SL, Salyer P, Grucza RA, Bierut LJ (2015) Twitter chatter about marijuana. J Adolesc Health 56:139–145. https://doi.org/10.1016/j.jadohealth.2014.10.270

Chang RM, Kauffman RJ, Kwon Y (2014) Understanding the paradigm shift to computational social science in the presence of big data. Decis Support Syst 63:67–80. https://doi.org/10.1016/j.dss.2013.08.008

Christen K, Merrill A, Wynne M (2017) A community of relations: Mukurtu hubs and spokes. D-Lib Magazine 23.https://doi.org/10.1045/may2017-christen

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 Netw 67:41–46. https://doi.org/10.1016/j.socnet.2019.10.010

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 Inf Technol. 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. J Commun 64:317–332. https://doi.org/10.1111/jcom.12084

Colombo GB, Burnap P, Hodorog A, Scourfield J (2016) Analysing the connectivity and communication of suicidal users on Twitter. Comput Commun 73:291–300. https://doi.org/10.1016/j.comcom.2015.07.018

Cooky C, Linabary JR, Corple DJ (2018) Navigating big data dilemmas: feminist holistic reflexivity in social media research. Big Data Soc 5:2053951718807731. https://doi.org/10.1177/2053951718807731

Croeser S, Highfield T (2020) Blended data: Critiquing and complementing social media datasets, big and small. In: Hunsinger J, Allen MM, Klastrup L (eds) Second international handbook of internet research. Springer, Netherlands, Dordrecht, pp 669–690

DDI Alliance (2022) Data Documentation Initiative. https://web.archive.org/web/20220202185335/https://ddialliance.org/

DocNow (2020) Documenting the Now. https://web.archive.org/web/20220419155938/https://www.docnow.io/. Accessed 22 Feb 2020

Doft D (2015) Facebook, Twitter, and the Wild West of IP enforcement on social media: weighing the merits of a uniform dispute resolution policy. J Marshall L Rev 49:959

Dolin-Mescal A (2018) Social humans. https://web.archive.org/web/20220208021334/https://www.docnow.io/social-humans/

Driscoll K, Walker S (2014) Working within a black box: transparency in the collection and production of big Twitter data. Int J Commun 8:20

EFF (2014) Unintended consequences: sixteen years under the DMCA. In: Electronic frontier foundation. https://web.archive.org/web/20220702055813/https://www.eff.org/wp/unintended-consequences-16-years-under-dmca

Ellison N, Heino R, Gibbs J (2006) Managing impressions online: self-presentation processes in the online dating environment. J Comput-Mediat Commun 11:415–441. https://doi.org/10.1111/j.1083-6101.2006.00020.x

Fan W, Gordon MD (2014) The power of social media analytics. Commun ACM 57:74–81. https://doi.org/10.1145/2602574

Fiesler C, Dye M, Feuston JL, Hiruncharoenvate C, Hutto CJ, Morrison S, Khanipour Roshan P, Pavalanathan U, Bruckman AS, De Choudhury M, Gilbert E (2017) What (or who) is public? Privacy settings and social media content sharing. In: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing. ACM, Portland, OR, pp 567–580

Fiesler C, Proferes N (2018) "Participant" perceptions of Twitter research ethics. Social Media + Society 4:205630511876336. https://doi.org/10.1177/2056305118763366

Franzke AS, Bechmann A, Ess CM, Zimmer M (2020) Internet research: ethical guidelines 3.0. AoIR (The International Association of Internet Researchers)

Ghermandi A, Sinclair M (2019) Passive crowdsourcing of social media in environmental research: a systematic map. Glob Environ Chang 55:36–47. https://doi.org/10.1016/j.gloenvcha.2019.02.003

Goffman E (1959) The presentation of self in everyday life. Doubleday Anchor Books, New York, NY

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. J Med Internet Res 21:e7081. https://doi.org/10.2196/jmir.7081

González-Bailón S (2013) Social science in the era of big data. Policy Internet 5:147–160. https://doi.org/10.1002/1944-2866.POI328

Greene T, Shmueli G, Ray S, Fell J (2019) Adjusting to the GDPR: the impact on data scientists and behavioral researchers. Big Data 7:140–162. https://doi.org/10.1089/big.2018.0176

Halavais A (2015) Bigger sociological imaginations: framing big social data theory and methods. Inf Commun Soc 18:583–594. https://doi.org/10.1080/1369118X.2015.1008543

Hargittai E (2020) Potential biases in big data: omitted voices on social media. Soc Sci Comput Rev 38:10–24. https://doi.org/10.1177/0894439318788322

Hemphill L, Leonard SH, Hedstrom M (2018) Developing a social media archive at ICPSR. In: Web Archiving and Digital Libraries (WADL). Fort Worth, TX

Henderson KA, Spinello RA, Lipinski TA (2007) Prudent policy? Reassessing the digital millennium copyright act. SIGCAS Comput Soc 37:25–40. https://doi.org/10.1145/1327325.1327327

Hogan B (2010) The presentation of self in the age of social media: distinguishing performances and exhibitions online. Bull Sci Technol Soc 30:377–386. https://doi.org/10.1177/0270467610385893

Hökby S, Hadlaczky G, Westerlund J, Wasserman D, Balazs J, Germanavicius A, Machín N, Meszaros G, Sarchiapone M, Värnik A, Varnik P, Westerlund M, Carli V (2016) Are mental health effects of internet use attributable to the web-based content or perceived consequences of usage? A longitudinal study of European adolescents. JMIR Mental Health 3:e31. https://doi.org/10.2196/mental.5925

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, London, UK

Hutton L, Henderson T (2013) An architecture for ethical and privacy-sensitive social network experiments. SIGMETRICS Perform Evaluat Rev 40:90–95. https://doi.org/10.1145/2479942.2479954

Ito M (2008) Introduction. In: Varnelis K (ed) Networked publics. MIT Press, Cambridge, MA, pp 1–14

Jendryke M, Balz T, McClure SC, 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. Comput Environ Urban Syst 62:99–112. https://doi.org/10.1016/j.compenvurbsys.2016.10.004

Jules B, Summers E, Mitchell VJr (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

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 Soc, 4.https://doi.org/10.1177/2053951717736336

King G, Persily N (2020) A new model for industry–academic partnerships. PS: Polit Sci Polit 53:703–709. https://doi.org/10.1017/S1049096519001021

Kirkegaard EOW, Bjerrekær JD (2016) The OKCupid dataset: a very large public dataset of dating site users. Open Differ Psychol

Kitchin R (2014) Big data, new epistemologies and paradigm shifts. Big Data Soc, 1.https://doi.org/10.1177/2053951714528481

Kramer ADI, Guillory JE, Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. Proc Natl Acad Sci 111:8788–8790. https://doi.org/10.1073/pnas.1320040111

Lanchester J (2017) You are the product. London Review of Books 39

Latour B (2007) Beware, your imagination leaves digital traces. Times Higher Literary Suppl 6:129–131

Lazer D, Pentland A, Adamic L, Aral S, Barabasi A-L, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Van Alstyne M (2009) Computational social science. Science 323:721–723. https://doi.org/10.1126/science.1167742

Lee TB (2006) Circumventing competition: the perverse consequences of the Digital Millennium Copyright Act. Policy Analysis

Lewis K, Kaufman J, Gonzalez M, Wimmer A, Christakis N (2008) Tastes, ties, and time: a new social network dataset using Facebook.com. Soc Netw 30:330–342. https://doi.org/10.1016/j.socnet.2008.07.002

Lorentzen DG, Nolin J (2017) Approaching completeness: capturing a hashtagged Twitter conversation and its follow-on conversation. Soc Sci Comput Rev 35:277–286. https://doi.org/10.1177/0894439315607018

Madden M (2014) Public perceptions of privacy and security in the post-Snowden era. Pew Research Center

Mannheimer S, Hull EA (2018) Sharing selves: developing an ethical framework for curating social media data. Int J Digit Curation 12:196–209. https://doi.org/10.2218/ijdc.v12i2.518
Manovich L (2012) Trending: the promises and the challenges of big social data. In: Gold MK (ed) Debates in the digital humanities. University of Minnesota Press, Minneapolis, MN, pp 460–475
Markham A (2012) Fabrication as ethical practice. Inf Commun Soc 15:334–353. 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. Comput Environ Urban Syst 74:161–174. https://doi.org/10.1016/j.compenvurbsys.2018.11.001
Marwick AE, Boyd D (2014) Networked privacy: how teenagers negotiate context in social media. New Media Soc 16:1051–1067. https://doi.org/10.1177/1461444814543995
Marwick AE, Boyd D (2011) I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. New Media Soc 13:114–133.https://doi.org/10.1177/1461444810365313
Mathur A, Bleckman JD, Lyle J (2017) Reuse of restricted-use research data. Curating research data, volume two: a handbook of current practice. Association of College and Research Libraries, Chicago, IL, pp 258–261
McCook AA (2016) Publicly available data on thousands of OKCupid users pulled over copyright claim. In: Retraction watch. https://retractionwatch.com/2016/05/16/publicly-available-data-on-thousands-of-okcupid-users-pulled-over-copyright-claim/. Accessed 28 Apr 2022
McKee R (2013) Ethical issues in using social media for health and health care research. Health Policy 110:298–301. https://doi.org/10.1016/j.healthpol.2013.02.006
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 J Corporate Financ Commer Law 16:279
Metcalf J (2016) Big data analytics and revision of the common rule. Commun ACM 59:31–33. 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 Soc, 3.https://doi.org/10.1177/2053951716650211
Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C (2013) A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 15:e85. https://doi.org/10.2196/jmir.1933
Moreno JL (1934) Who shall survive?: A new approach to the problem of human interrelations. Nervous and Mental Disease Publishing Co, Washington, DC
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
Neale B, Bishop L (2012) The Timescapes Archive: a stakeholder approach to archiving qualitative longitudinal data. Qual Res 12:53–65. https://doi.org/10.1177/1468794111426233
Nebeker C, Dunseath SE, Linares-Orozco R (2020) A retrospective analysis of NIH-funded digital health research using social media platforms. Digital Health.https://doi.org/10.1177/2055207619901085
Nippert-Eng CE (2010) Islands of privacy. The University of Chicago Press, Chicago, IL
Nissenbaum H (2009) Privacy in context: technology, policy, and the integrity of social life. Stanford University Press, Palo Alto, CA
Obar JA, Oeldorf-Hirsch A (2020) The biggest lie on the internet: ignoring the privacy policies and terms of service policies of social networking services. Inf Commun Soc 23:128–147. 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.https://doi.org/10.5210/fm.v17i7.3993

Osterberg G (2017) Update on the Twitter archive at the Library of Congress. In: 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/

Palen L, Dourish P (2003) Unpacking "privacy" for a networked world. In: Proceedings of the SIGCHI conference on human factors in computing systems. association for computing machinery, Ft. Lauderdale, Florida, USA, pp 129–136

Pasquetto IV, Borgman CL, Wofford MF (2019) Uses and reuses of scientific data: the data creators' advantage. Harvard Data Sci Rev, 1. https://doi.org/10.1162/99608f92.fc14bf2d

Paul M, Dredze M (2011) You are what you tweet: analyzing twitter for public health. Proceedings of the international AAAI conference on web and social media 5:265–272. https://doi.org/10.1609/icwsm.v5i1.14137

Petronio SS (2002) Boundaries of privacy: dialectics of disclosure. State University of New York Press, Albany, NY

Proferes N (2017) Reaction to Cornelius Puschmann. In: Kinder-Kurlanda K, Zimmer M (eds) Internet research ethics for the social age. Peter Lang, New York, NY, p 114

Prom CJ (2017) Social feed manager guide for building social media archives. University of Illinois at Urbana-Champaign

Puschmann C (2017) Bad judgment, bad ethics? Validity in computational social media research. In: Zimmer M, Kinder-Kurlanda K (eds) Internet research ethics for the social age. Peter Lang, New York, NY, pp 95–113

Puschmann C (2019) An end to the Wild West of social media research: a response to Axel Bruns. Inf Commun Soc 22:1582–1589. https://doi.org/10.1080/1369118X.2019.1646300

Puschmann C, Burgess J (2014) The politics of Twitter data. In: Weller K, Bruns A, Burgess J, Puschmann C, Mahrt M (eds) Twitter and society. Peter Lang, New York, NY, pp 43–54

Rains SA, Brunner SR (2015) What can we learn about social network sites by studying Facebook? A call and recommendations for research on social network sites. New Media Soc 17:114–131. https://doi.org/10.1177/1461444814546481

Reuter K, Zhu Y, Angyan P, Le N, Merchant AA, Zimmer M (2019) Public concern about monitoring Twitter users and their conversations to recruit for clinical trials: survey study. J Med Internet Res 21:e15455. https://doi.org/10.2196/15455

Rivers CM, Lewis BL (2014) Ethical research standards in a world of big data. F1000 Research 3:38. https://doi.org/10.12688/f1000research.3-38.v2

Ruthven I, Buchanan S, Jardine C (2018) Relationships, environment, health and development: the information needs expressed online by young first-time mothers. J Am Soc Inf Sci 69:985–995. https://doi.org/10.1002/asi.24024

Salganik MJ (2018) Bit by bit: social research in the digital age. Princeton University Press, Princeton, NJ

Schema.org (2020) Data and datasets. https://web.archive.org/web/20211215014211/https://schema.org/docs/data-and-datasets.html

Schneble CO, Elger BS, Shaw D (2018) The Cambridge Analytica affair and internet-mediated research. EMBO reports 19:e46579. https://doi.org/10.15252/embr.201846579

Secretary's Advisory Committee on Human Research Protections (2013) Considerations and recommendations concerning internet research and human subjects research regulations, with revisions

Secretary's Advisory Committee on Human Research Protections (2015) Attachment A: human subjects research implications of "big data"

Segerberg A, Bennett WL (2011) Social media and the organization of collective action: using Twitter to explore the ecologies of two climate change protests. Commun Rev 14:197–215. https://doi.org/10.1080/10714421.2011.597250

Shah DV, Cappella JN, Neuman WR (2015) Big data, digital media, and computational social science: possibilities and perils. Ann Am Acad Pol Soc Sci 659:6–13. https://doi.org/10.1177/0002716215572084

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. In: 49th Hawaii international conference on system sciences (HICSS). IEEE, Koloa, HI, pp 1909–1918

Sieber JE (1991) Sharing social science data: advantages and challenges. Sage Publications, Thousand Oaks, CA

Simmel G (1955) Conflict and the web of group affiliations. The Free Press, New York, NY

Sleeper M, Balebako R, Das S, McConahy AL, Wiese J, Cranor LF (2013) The post that wasn't: exploring self-censorship on Facebook. In: Proceedings of the 2013 conference on computer supported cooperative work—CSCW '13. ACM Press, San Antonio, Texas, USA, p 793

Sloan L (2016) Social science 'lite'? Deriving demographic proxies from Twitter. The Sage handbook of social media research methods. Sage Publications, London, UK, pp 90–104

Stier S, Breuer J, Siegers P, Thorson K (2020) Integrating survey data and digital trace data: Key issues in developing an emerging field. Soc Sci Comput Rev 38:503–516. https://doi.org/10.1177/0894439319843669

Stoycheff E, Liu J, Wibowo KA, Nanni DP (2017) What have we learned about social media by studying Facebook? A decade in review. New Media Soc 19:968–980. https://doi.org/10.1177/1461444817695745

Sujon Z (2017) Reaction to tromble and stockmann. In: Kinder-Kurlanda K, Zimmer M (eds) Internet research ethics for the social age. Peter Lang, New York, NY

Taylor J, Pagliari C (2018) Mining social media data: how are research sponsors and researchers addressing the ethical challenges? Research Ethics 14:1–39. https://doi.org/10.1177/1747016117738559

The Economist (2022) Your secret's safe with me; Data privacy. The Economist, 62–63

Thomson SD (2016) Preserving social media. Digital Preservation Coalition Technology Watch Report.https://doi.org/10.7207/twr16-01

Törnberg P, Törnberg A (2018) The limits of computation: a philosophical critique of contemporary big data research. Big Data Soc 5:2053951718811843. https://doi.org/10.1177/2053951718811843

Twitter (2023a) Compliance Firehouse API: honoring user intent on Twitter. In: Twitter developer platform. https://web.archive.org/web/20230329190341/https://developer.twitter.com/en/docs/twitter-api/enterprise/compliance-firehose-api/guides/honoring-user-intent

Twitter (2023b) Developer policy: content redistribution. https://web.archive.org/web/20230403102334/https://developer.twitter.com/en/developer-terms/policy

Twitter (2023c) Developer terms: more about restricted uses of the Twitter APIs. https://web.archive.org/web/20230401142538/https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

UK Central Digital and Data Office (2020) Data ethics framework. In: Gov.uk. https://web.archive.org/web/20230310054327/https://www.gov.uk/government/publications/data-ethics-framework

U.S. Department of Health and Human Services (1991) Federal policy for the protection of human subjects ("Common rule"). HHS.gov

Varol O, Ferrara E, Davis CA, Menczer F, Flammini A (2017) Online human-bot interactions: detection, estimation, and characterization. In: Proceedings of the eleventh international AAAI conference on web and social media (ICWSM 2017). AAAI Publications, Montreal, Canada, p 10

Verma IM (2014) Editorial expression of concern: experimental evidence of massivescale emotional contagion through social networks. Proc Natl Acad Sci 111:10779–10779. 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 10:62. https://doi.org/10.3390/fi10070062

Villarroel Ordenes F, Grewal D, Ludwig S, Ruyter KD, Mahr D, Wetzels M (2019) Cutting through content clutter: how speech and image acts drive consumer sharing of social media brand messages. J Consumer Res 45:988–1012. 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. Int J Digital Human 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, Cham, Switzerland

Washington Post (2018) Transcript of Mark Zuckerberg's Senate hearing

Weller K, Kinder-Kurlanda KE (2016) A manifesto for data sharing in social media research. In: Proceedings of the 8th ACM conference on web science—WebSci '16. ACM Press, Hannover, Germany, pp 166–172

Wilkinson D, Thelwall M (2011) Researching personal information on the public web: methods and ethics. Soc Sci Comput Rev 29:387–401. https://doi.org/10.1177/0894439310378979

Wilkof N (2016) IP knowledge in the age of Wikipedia and the blogosphere. J Intellect Property Law Pract 11:477–478. https://doi.org/10.1093/jiplp/jpw072

Williams SA, Terras MM, Warwick C (2013) What do people study when they study Twitter? Classifying Twitter related academic papers. J Documen 69:384–410. https://doi.org/10.1108/JD-03-2012-0027

Wilson RE, Gosling SD, Graham LT (2012) A review of Facebook research in the social sciences. Perspect Psychol Sci 7:203–220. https://doi.org/10.1177/1745691612442904

Wittwer M, Reinhold O, Alt R (2017) Capturing customer context from social media: mapping social media API and CRM profile data. In: Proceedings of the international conference on web intelligence. Association for computing machinery, leipzig, Germany, pp 993–997

Zhang Z, He Q, Gao J, Ni M (2018) A deep learning approach for detecting traffic accidents from social media data. Transp Res Part C: Emerg Technol 86:580–596. https://doi.org/10.1016/j.trc.2017.11.027

Zimmer M (2018) Addressing conceptual gaps in big data research ethics: an application of contextual integrity. Social Media + Society 4. https://doi.org/10.1177/2056305118768300

Zimmer M (2016) OkCupid study reveals the perils of big-data science. Wired

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 (2010) "But the data is already public": On the ethics of research in Facebook. Ethics Inf Technol 12:313–325. https://doi.org/10.1007/s10676-010-9227-5

Zimmer M, Proferes NJ (2014) A topology of Twitter research: Disciplines, methods, and ethics. Aslib J Inf Manag 66:250–261. https://doi.org/10.1108/AJIM-09-2013-0083

Comparison of Issues and Data Curation Strategies

5

The literature reviewed in Chaps. 3 and 4 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 issues of context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. 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 rarely been conducted on a large scale.

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). In this chapter, I synthesize the literature reviewed in Chaps. 3 and 4, comparing key issues relating to qualitative data reuse and big social research, and highlighting data curation practices that support epistemologically sound, ethical, and legal use and reuse of qualitative and big social data.[1]

[1] Parts of this chapter were originally published in: Mannheimer (2021) Data curation implications of qualitative data reuse and big social research. Journal of eScience Librarianship 10:e1218. https://doi.org/10.7191/jeslib.2021.1218.

S. Mannheimer, *Scaling Up: How Data Curation Can Help Address Key Issues in Qualitative Data Reuse and Big Social Research*, Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-49222-8_5

5.1 Context

For both qualitative data reuse and big social research, there is concern that data may be misconstrued or misused if they are shared without sufficient contextual details. For qualitative data reuse, 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 research, the problem is that individual posts are removed from their context by the very nature of the research process itself, which, through large-scale data collection, isolates posts from the larger context of data creator's personal and public life. During the big social research process, the researcher may never speak to the people who created the posts and may never know these creators' identities or broader contexts. As mentioned in Chap. 4, Sect. 4.3.1, Marwick and boyd (2011) also refer to a "context collapse" in big social data, suggesting that "multiple audiences [flatten] into one" when posting on social media, 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 qualitative data: when researchers share qualitative data, the future audience and use cases of those data are unknown.

For both big social research and qualitative data reuse, the literature suggests that the full context and meaning of data may never be accurately understood by qualitative data reusers or big social researchers. However, using data curation strategies to communicate as much context as possible can help support meaningful data use and reuse.

5.1.1 Data Curation for Context

As summarized above, understanding data outside of its original context is difficult for both qualitative data reuse and big social research. However, data curation strategies can, to some extent, help communicate contextual details for shared data, thus supporting meaningful data use and reuse.

Clear documentation

- For qualitative data: Data curators can encourage contextual documentation throughout the research process, to be published alongside qualitative data. This could include documentation about research methods and practices, consent form, IRB approval numbers 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).
- For big social research: Data curators can encourage as much documentation as possible of the methods, communities, and platforms. Context can also be communicated through metadata such as geolocation, @-mentions, or hashtags.

- Initiatives such as Annotation for Transparent Inquiry (Karcher and Weber 2019), Open Context (Kansa and Kansa 2018), and the Data Curation Network (Johnston et al. 2018) all support researchers and data repositories in creating documentation to encourage contextual integrity for data reuse.

Archiving related data

- Repositories may also choose to archive (or link to archived versions of) web URLs, images, and other resources (Thomson 2016).

5.2 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 often relate to 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 Sect. 5.3. 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 compound.

5.2.1 Data Curation for Data Quality and Trustworthiness

While data curation is not a simple solution to these challenges, clear documentation, use of trustworthy repositories, and linking to related datasets are all discussed in the literature as strategies to support data quality and trustworthiness.

Clear documentation

- Data curators can support documentation of the research process when sharing data, including documenting any potential bias, errors, or missing data.
- Data curators can conduct quality checks on descriptive metadata.

Trustworthy repositories

- Data repositories and academic libraries can contribute to data quality and trustworthiness by supporting data management, curation, and metadata (Giarlo 2013; Frank et al. 2017; Yoon and Lee 2019).
- Trust in data can be enhanced by trust in the repository where it is archived. To support healthy infrastructure and long-term preservation for repositories, initiatives such as the CoreTrustSeal Trustworthy Data Repositories Requirements (CoreTrustSeal 2023) and the Audit and Certification of Trustworthy Digital Repositories (Consultative Committee on Space Data Systems 2011) provide community standards for repositories.
- Repositories and curators can also refer to the TRUST Principles, which complement the FAIR Principles to support trustworthy data stewardship for archived data (Lin et al. 2020).

Related and combined datasets

- Some researchers have attempted to create more representative datasets by blending big social data with smaller social datasets, or by combining multiple shared qualitative datasets. This strategy helps incorporate a broader range of perspectives than are present in a single dataset (Croeser and Highfield 2020). Data curators could provide links between related datasets to support future use. However, combining datasets comes with its own set of challenges (see Sect. 5.3, Data comparability, below).

5.3 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. For big social data, comparability is additionally affected by 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 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.

5.3.1 Data Curation for Data Comparability

The literature suggests that data curators can support data comparability by helping researchers create clear documentation, and by advocating for interoperable metadata standards.

Clear documentation

- For both qualitative data reuse and big social research, data curators can support comparability by encouraging researchers who publish data to include clear documentation to address missing data, research questions, and methods.

Metadata standards

- For both types of data, data curators can adapt existing standards such as DDI (DDI Alliance 2022) and Qualitative Data Exchange Schema (Corti and Gregory 2011) 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.
- The research and data curation communities can advocate for interoperable metadata standards that can be adopted by social media platforms and other big social data sources.

5.4 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 specifications for the use of 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 are reused—future use is difficult to predict. 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.

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 difficult 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 Secretary's Advisory Committee on Human Research Protections 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 Sect. 5.5. Privacy and confidentiality, for further discussion of identifiability).

5.4.1 Data Curation for Informed Consent

To mitigate some challenges of informed consent, data curators may be able to provide guidance on alternative consent strategies, as outlined below. Note that these strategies rely on data curators connecting with researchers before the research process begins—a challenge that is discussed further in Chap. 7, Sect. 7.3.1. Planning ahead for data curation.

Alternative consent strategies for qualitative data reuse

- If data curators can connect with researchers early in the research process, they can help researchers draft broad consent language to support data reuse (Kirilova and Karcher 2017).

- Researchers, curators, and IRBs can also work together to support tiered consent models, allowing research participants to select the level of data sharing with which they are comfortable.

Alternative consent strategies for big social research

- If data curators can connect with researchers early in the research process, they can encourage strategies such as focus groups, community advisory boards, or software-supported strategies for obtaining individual informed consent within social media platforms.

5.5 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). Many 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.

5.5.1 Data Curation for Privacy and Confidentiality

To address some of the privacy challenges reviewed above, data curation and data repository services have been developed to provide deidentification support, restricted data access, and data use agreements.

De-identification procedures

- 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.

Restricted access

- 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

- Data curators and repositories can provide customizable data use agreements that dictate the conditions required for other researchers to access and reuse the data. The data use agreement includes terms that the user must agree to follow if they download the data. For example, the agreement may stipulate that the data be used for academic research purposes, that the research be approved by an institutional review board, or that the researcher not attempt to reidentify the data (ICPSR 2018; QDR 2019).

5.6 Intellectual Property and Data Ownership

As discussed in Chap. 3, qualitative data may be the intellectual property of the research participants, who would need to either waive their rights or license their responses for use in the research study. Additionally, universities may claim ownership of academic research data. Consent agreements can clarify intellectual property and data ownership by outlining the rights of participants and institutions, as well as the responsibilities and obligations of future researchers using the data. The doctrine of fair use may also apply to qualitative data, since research that reuses data 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. Many repositories offer licensing options or suggest a Creative Commons Attribution (CC-BY) license or an Open Data Commons Attribution (ODC-By) license, and some in the data curation community encourage releasing data into the public domain using the Creative Commons public domain waiver (CC0) or Open Data Commons public domain dedication and license (ODC-PDDL) to simplify data reuse and rights management (Schofield et al. 2009; Schaeffer 2011).

Big social data sharing is made more complex by the fact that big social 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 Chap. 4, Sect. 4.3.6). 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).

5.6.1 Data Curation for Intellectual Property and Data Ownership

Data curators can support intellectual property challenges through rights management guidance, data licensing, and alternative archiving strategies.

Rights management for both big social research and qualitative data reuse

- Data curators and data repositories can help researchers with rights management—understanding how they can and cannot reuse shared data.
- For big social research, data curators can help researchers navigate terms of service to collect, archive, and share data in accordance with these terms.

Data licensing for qualitative data

- For qualitative data, if data curators can reach researchers early in the process, they can ensure that data licensing language is included as part of initial consent agreements.
- Data curators have another opportunity to discuss data licensing at the point of data archiving and sharing.

Alternative archiving strategies for big social data

- If raw data cannot be archived, data repositories can archive associated information such as data workflows and code that can allow future users to replicate the data collection and analysis process (Hemphill et al. 2018).
- Data repositories maybe able to archive representative metadata such as lists of TweetIDs.
- Data curators can encourage inclusion of tools such as the Twitter Hydrator as part of the data deposit, to support usability for the archived data (Kinder-Kurlanda et al. 2017).

5.7 Summary of Similarities and Differences

The issues described above are all key to successful qualitative data reuse and big social research. However, 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 and Wolf 2010; Verma 2014). 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 explored further in Chaps. 6 and 7, which describe new insights derived from semi-structured interviews with researchers and data curators.

References

Consultative Committee on Space Data Systems (2011) Audit and certification of trustworthy digital repositories: recommended practice (CCSDS 652.0-M-1). Consultative Committee on Space Data Systems, Washington, DC

CoreTrustSeal (2023) Core trustworthy data repositories requirements. https://web.archive.org/web/20230407041240/https://www.coretrustseal.org/

Corti L, Gregory A (2011) CAQDAS comparability: what about CAQDAS data exchange? Forum Qual Sozialforschung/Forum Qualitat Soc Res, 12. https://doi.org/10.17169/FQS-12.1.1634

Croeser S, Highfield T (2020) Blended data: critiquing and complementing social media datasets, big and small. In: Hunsinger J, Allen MM, Klastrup L (eds) Second international handbook of internet research. Springer, Netherlands, Dordrecht, pp 669–690

DDI Alliance (2022) Data Documentation Initiative. https://web.archive.org/web/20220202185335/https://ddialliance.org/

Frank RD, Chen Z, Crawford E, Suzuka K, Yakel E (2017) Trust in qualitative data repositories. Proc Assoc Inf Sci Technol 54:102–111. https://doi.org/10.1002/pra2.2017.14505401012

Giarlo M (2013) Academic libraries as data quality hubs. J Librarianship Scholarly Commun 1:eP1059. https://doi.org/10.7710/2162-3309.1059

Hemphill L, Leonard SH, Hedstrom M (2018) Developing a social media archive at ICPSR. In: Web Archiving and Digital Libraries (WADL). Fort Worth, TX

ICPSR (2012) Guide to social science data preparation and archiving: introduction. http://www.icpsr.umich.edu/files/deposit/dataprep.pdf

ICPSR (2018) Restricted data use agreement for restricted data from the Inter-University Consortium for Political and Social Research (ICPSR). Inter-university Consortium for Political and Social Research (ICPSR), University of Michigan

Johnston LR, Carlson J, Hudson-Vitale C, Imker H, Kozlowski W, Olendorf R, Stewart C, Blake M, Herndon J, McGeary TM, Hull E (2018) Data Curation Network: a cross-institutional staffing

model for curating research data. Int J Digit Curation 13:125–140. https://doi.org/10.2218/ijdc.v13i1.616

Kansa SW, Kansa EC (2018) Data beyond the archive in digital archaeology: an introduction to the special section. Adv Archaeol Pract 6:89–92. https://doi.org/10.1017/aap.2018.7

Karcher S, Weber N (2019) Annotation for transparent inquiry: transparent data and analysis for qualitative research. IASSIST Quarterly 43:1–9. https://doi.org/10.29173/iq959

Kirilova D, Karcher S (2017) Rethinking data sharing and human participant protection in social science research: applications from the qualitative realm. Data Sci J 16:43. https://doi.org/10.5334/dsj-2017-043

Lin D, Crabtree J, Dillo I, Downs RR, Edmunds R, Giaretta D, De Giusti M, L'Hours H, Hugo W, Jenkyns R, Khodiyar V, Martone ME, Mokrane M, Navale V, Petters J, Sierman B, Sokolova DV, Stockhause M, Westbrook J (2020) The TRUST Principles for digital repositories. Scientific Data 7:144. https://doi.org/10.1038/s41597-020-0486-7

Mannheimer S (2021) Data curation implications of qualitative data reuse and big social research. J eSci Librariansh 10:e1218. https://doi.org/10.7191/jeslib.2021.1218

Marwick AE, boyd d, (2011) I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. New Media Soc 13:114–133. https://doi.org/10.1177/1461444810365313

Mello MM, Wolf LE (2010) The Havasupai Indian Tribe case—lessons for research involving stored biologic samples. N Engl J Med 363:204–207. https://doi.org/10.1056/NEJMp1005203

QDR (2019) General terms and conditions of use. Syracuse University, The Qualitative Data Repository

Schaeffer P (2011) Why does Dryad use CC0? In: Dryad news and views. https://web.archive.org/web/20220401073936/https://blog.datadryad.org/2011/10/05/why-does-dryad-use-cc0/

Schofield PN, Bubela T, Weaver T, Portilla L, Brown SD, Hancock JM, Einhorn D, Tocchini-Valentini G, Hrabe de Angelis M, Rosenthal N (2009) Post-publication sharing of data and tools. Nature 461:171–173. https://doi.org/10.1038/461171a

Sujon Z (2017) Reaction to Tromble and Stockmann. In: Kinder-Kurlanda K, Zimmer M (eds) Internet research ethics for the social age. Peter Lang, New York, NY

Summers E (2017) The catalog and the hydrator. In: Documenting the Now. https://news.docnow.io/the-catalog-and-the-hydrator-3299eddfe21e

Thomson SD (2016) Preserving social media. Digital Preservation Coalition Technology Watch Report.https://doi.org/10.7207/twr16-01

Verma IM (2014) Editorial expression of concern: experimental evidence of massivescale emotional contagion through social networks. Proc Natl Acad Sci 111:10779–10779. https://doi.org/10.1073/pnas.1412469111

Yoon A, Lee YY (2019) Factors of trust in data reuse. Online Information Review. https://doi.org/10.1108/OIR-01-2019-0014

Zimmer M (2010) "But the data is already public": On the ethics of research in Facebook. Ethics Inf Technol 12:313–325. https://doi.org/10.1007/s10676-010-9227-5

Researchers and Data Curators Respond to Key Issues

6

Up until this point in the book, I have reviewed and examined existing literature to synthesize key issues. In this chapter and in Chap. 7, I will discuss the results of semi-structured interviews with participants from three communities of practice: qualitative researchers who have shared or reused data (for simplicity, I call these participants *qualitative researchers* throughout the rest of the chapter), big social researchers, and data curators. By speaking directly to participants about their experiences and concerns, I aim to build conclusions about the similarities and differences in how each community of practice addresses the issues of context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. In this chapter, I also identify three new themes relevant to qualitative data reuse, big social research, and data curation: domain differences, strategies for responsible practice, and perspectives on data curation and data sharing.

This chapter provides detailed interview results, including quotes that illustrate participants' ideas in their own words. The reader may choose to read this chapter in its entirety, or they may selectively read the sections most relevant to their interests. Chapter 7 synthesizes the results of the interviews with researchers and data curators into key insights.

S. Mannheimer, *Scaling Up: How Data Curation Can Help Address Key Issues in Qualitative Data Reuse and Big Social Research*, Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-49222-8_6

Table 6.1 Qualitative researchers by discipline

Discipline	Number of participants
Information science	4
Anthropology	2
Public health	1
Education	1
Nursing	1
Social work	1

Table 6.2 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

6.1 A Brief Overview of Participants and Methods[1]

I conducted interviews with thirty participants—ten qualitative researchers, ten big social researchers, and ten data curators, all of whom work in the United States. The qualitative researchers and big social researchers whom I interviewed came from a variety of disciplines. Qualitative researchers and big social researchers from the discipline of Information Science are somewhat over-represented in my dataset because Information Science researchers were more likely to respond positively to my interview requests. Because Information Science leans toward interdisciplinarity (Chang 2018), the the different examples discussed by Information Science researchers were distinct enough that the sample still provides a broad variety of disciplinary ideas. Tables 6.1 and 6.2 provide overviews of the disciplines of the qualitative and big social researchers interviewed for this book.

Participants also 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 6.3 provides an overview of the number of participants from each rank or role.

[1] For full sampling information, interview transcripts, content analysis, and codebook, please see the associated dataset in Qualitative Data Repository (Mannheimer 2023).

Table 6.3 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

In the interviews, I asked participants to focus on a specific incident in which they shared or reused qualitative data, conducted big social research, or curated qualitative or big social data data. I conducted a qualitative content analysis of the interview transcripts, using a combination of deductive and inductive coding approaches. The interviews were structured according to the six key issues identified from the literature (see Chaps. 3 and 4)—context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. I deductively coded a parent theme for each of these key issues. I then used inductive coding to create subthemes beneath each of the parent themes.

6.2 Interview Results

Participants from each community of practice (big social researchers, qualitative researchers, and data curators) discussed each of the six key issues previously identified:

- Context
- Data quality and trustworthiness
- Data comparability
- Informed consent
- Privacy and confidentiality
- Intellectual property and data ownership.

Additionally, the interviews revealed an additional three important themes:

- Domain differences
- Strategies for responsible practice
- Perspectives on data curation and sharing.

These three additional themes proved to be analytically powerful lenses through which the participants viewed big social research, qualitative data reuse, and data curation. These themes explore how each community of practice understands their own disciplinary and methodological foundations and landscapes, the strategies that each community of practice used to support responsible practice in their own research, and each community's perspectives on data curation and data sharing.

I organize the results below by issue, with subsections that specifically discuss interview themes from each community of practice—qualitative researchers, big social researchers, and data curators. 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. 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. 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. I use *n* to indicate numbers within the full set of thirty participants. When numbers appear within a community of practice subsection, I use either *qr* (for qualitative researchers), *bsr* (for big social researchers), or *dc* (for data curators), to indicate that these 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.

6.2.1 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: First the process of creating thorough documentation is time-consuming (n = 7; qr = 5, bsr = 1, dc = 1). Second, too much contextual information may sacrifice participant privacy—that is, the greater amount of detailed contextual documentation is added to the data, the

more likely that individual participants' data will be identifiable (n = 10; qr = 4, bsr = 0, dc = 6). The theme of privacy and confidentiality is discussed further in Sect. 6.2.5.

Most participants had considered the idea of context, but different communities or practice (big social researchers, qualitative researchers, and data curators) spoke about different central concerns regarding context and different strategies for preserving the context of data. Qualitative researchers were most concerned with strategies that could be used by the original researcher to communicate the context of shared data. Big social researchers were most concerned with technical considerations around context, and data curators were most concerned with contextual documentation, description, and metadata. I provide more detail on each community of practice below.

6.2.1.1 Context—Qualitative Researchers

The qualitative researchers I interviewed tended to focus on how an original researcher could communicate the context of shared data. Qualitative researchers (qr = 4) discussed how their own methods, ideas, and values as researchers contributed to the context of the data. Some (qr = 2) suggested collaborating with original researchers to incorporate their contextual expertise into a data reuse project. Others (qr = 3) suggested that a degree of misinterpretation may be inevitable, and that context could never be fully communicated; however, 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. 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 (qr = 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 [U.S. State] eating disorder group or something along those lines. That bothers me because if it's central [U.S. State], that means it could be urban. It could be [City] where I live right now, which is quite urban and Black and socio-economically divided. Or it could be central [U.S. 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)

These examples illustrate how qualitative researchers conducted informal risk–benefit analyses throughout their research and data sharing processes: 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 Sect. 6.2.8.

6.2.1.2 Context—Big Social Researchers

Big social researchers' discussion of context often focused on technical considerations. Some big social researchers (bsr = 4) talked about how data mining techniques remove the user interface as a contextual factor, leaving just text 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 to support clearer context (bsr = 3). For instance, one big social researcher described analyzing book-reading data from a social media platform in a similar way as the social media platform itself, 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 (bsr = 2) also talked about selecting data that had more inherent context, such as selecting Tweets that include geographical location tags.

Representativeness of the data was also a key topic for big social researchers (bsr = 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 big social data was also discussed in relation to data quality and trustworthiness (see Sect. 6.2.2).

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 Sect. 6.2.5. for further discussion of the theme of privacy and confidentiality.

6.2.1.3 Context—Data Curators

Data curators were most likely to talk about documentation, description, and metadata as a strategy for preserving context; seven data curators discussed this topic, compared to four qualitative researchers and two big social researchers. Data curators identified context as a key to understanding archived data (dc = 2), and they also emphasized the importance of preserving related materials alongside archived data (dc = 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,[2] hoping that will be around longer" (DC09). This focus on the value of digital preservation, in addition to sharing and reuse, was unique to data curators.

[2] Perma.cc is a web archiving service for legal and academic citations (Perma.cc 2023).

Like qualitative researchers, data curators discussed the tension between providing contextual information and preserving privacy for human subjects (dc = 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 the fullest picture that they can take away from qualitative datasets. (DC10)

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.

6.2.2 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 to facilitate and support data reuse. Similarly, all three communities indicated that they were more likely to find data trustworthy for their own use 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 the 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 qualitative researchers, big social researchers, and data curators. Rather, as further detailed below, each community of practice emphasized its own specialized considerations pertaining to data quality and trustworthiness.

6.2.2.1 Data Quality and Trustworthiness—Qualitative Researchers

Qualitative researchers usually included a discussion of quality in their manuscript (qr = 5), rather than describing quality in a readme or other via descriptive metadata that would be included alongside their published data. As one researcher explained, "[I wrote], '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)

Qualitative researchers emphasized the inherent messiness of conducting research with human participants. Some qualitative researchers discussed the quality difference when reusing secondary data as opposed to talking directly with research participants (qr = 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 (qr = 2). For example, one qualitative researcher said of 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 [between researcher and participant]" (QR04).

Lastly, qualitative researchers were aware of researcher bias and the ways in which the researchers themselves could affect the qualitative research process (qr = 3). As one interviewee who researches employee experiences at work 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).

6.2.2.2 Data Quality and Trustworthiness—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[ument] 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 Sect. 6.2.1.2, big social researchers spoke about the representativeness of social media data, highlighting that using a non-representative dataset affected data quality (bsr = 3). Big social researchers (bsr = 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 have 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 (bsr = 3), whereas others were aware of spam but decided that their research didn't necessitate removing it (bsr = 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 (bsr = 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 (bsr = 2); one researcher collected Reddit data using a third-party app in addition to using 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 (bsr = 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 that described looking to existing literature for guidance on data quality (bsr = 2), reading similar research to their own to see how data quality issues were addressed.

6.2.2.3 Data Quality and Trustworthiness—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 technical aspects of data quality—ensuring that data are readable, and facilitating high quality documentation, description, and metadata. As one data curator phrased it, a full "description of the process, I think, should enhance trust for secondary users [by letting them] 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 (dc = 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 [common 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 and repository services as elements that support data quality and trustworthiness (dc = 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 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).

6.2.3 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).

6.2.3.1 Data Comparability—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 (qr = 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).

Some qualitative researchers (qr = 2) believed that increased amounts of data could lead to better conclusions, 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 (qr = 2) also pointed out that the complexity of qualitative data could hinder comparability, 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). Another researcher similarly suggested: "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).

6.2.3.2 Data Comparability—Big Social Researchers

Big social researchers described how more data sources could lead to stronger conclusions (bsr = 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 [we] 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 sources, big social researchers also discussed the challenges of matching up different datasets (bsr = 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 (bsr = 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…, but either way, you will need help from humans or technology to assist" (BSR08). Another researcher 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).

6.2.3.3 Data Comparability—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 (dc = 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 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 (dc = 5), curators discussed how documentation can help support comparability. One data curator discussed the importance of

contextual information when synthesizing multiple qualitative datasets: "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 combining data can lead to better conclusions (dc = 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 secondary data, but we don't have, 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, saying that have more data makes it easier for researchers to "approach that research longitudinally, like pull that same data [from year to] year, because it's more straightforward to do so" (DC10).

6.2.4 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 carry fewer concerns 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 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 in the sections below.

6.2.4.1 Informed Consent—Qualitative Researchers

While the majority of qualitative researchers (qr = 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 to do the 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 (qr = 7). Qualitative researchers who were reusing data (either their own data or historical data) (qr = 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 (qr = 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 (qr = 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 [this repository], here's how it would look by going into [the repository], 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 repository]" (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 (qr = 5), and speculated that participants may not always understand the nuances of a consent form (qr = 4). Some qualitative researchers (qr = 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 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 (qr = 3). Many talked about developing their own personal strategies and goals for responsible practice; this idea is discussed further in Sect. 6.2.8.

6.2.4.2 Informed Consent—Big Social Researchers

Big social researchers generally looked to IRBs to provide an ethical stamp of approval for their research. Only one (bsr = 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 (bsr = 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 (bsr = 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 (bsr = 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 (bsr = 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 (bsr = 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 they 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 (bsr = 2).

6.2.4.3 Informed Consent—Data Curators

Data curators discussed IRBs to support consent (dc = 6). Some viewed IRB documentation as a stamp of approval, or a way to encourage transparency for shared data. As one curator suggested, "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. One curator explained the challenges of dealing with IRBs from different institutions in this way:

> IRBs are only of limited help here, because a lot of IRBs think once data are deidentified, they're no longer human participant data. And so they kind of wave their hand. 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 [data 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 … 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 (dc = 8). Curators described a common situation of being approached by qualitative researchers who wanted to publish data, but whose research consent form did not clearly indicate 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 were still shareable; these decisions often depended on the sensitivity of the data (dc = 3) and whether the data were completely deidentified (see Sect. 6.2.5. 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 (dc = 2).

Like qualitative researchers and big social researchers, data curators spoke about the conflict between "publicly available" data and participants' expectations of privacy (dc = 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 (dc = 3), and the concern that it is impossible to know how data might be used in the future (dc = 3)—an idea that calls into question whether consent to data sharing can ever truly be *informed.*

6.2.5 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).

6.2.5.1 Privacy and Confidentiality—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 (qr = 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 (qr = 3). For example, one researcher described requirements for future users: "[Future users would have 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 (qr = 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… We had an Excel spreadsheet that was password protected" (QR10).

Qualitative researchers also considered potential harms to participants that could result from data sharing (qr = 2), especially for sensitive data or data from vulnerable populations (qr = 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).

6.2.5.2 Privacy and Confidentiality—Big Social Researchers

Despite the fact that big social researchers generally did not consider informed consent necessary for their research (see Sect. 6.2.4.2. 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 (bsr = 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 mashed 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 for privacy (bsr = 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. 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 (bsr = 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 (bsr = 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).

6.2.5.3 Privacy and Confidentiality—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 (dc = 7), and discussed using different levels of care depending on the sensitivity of the data (dc = 4), including being more stringent about data security (dc = 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 (dc = 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 assisting researchers with deidentification (dc = 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 too to social media data—as opposed to [for example] survey 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 (dc = 4) also discussed participant expectations around privacy and confidentiality in qualitative data reuse and big social research. One data curator described participant expectations for shared qualitative data: "[Participants 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.[3]

> 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)

6.2.6 Intellectual Property and Data Ownership

Compared to other key issues identified in this book, issues relating to intellectual property and data ownership were less clearly understood by the participants. A common idea discussed in interviews was the participants' lack of clarity about intellectual property rights and data ownership (n = 5; qr = 1, bsr = 2, dc = 2). Members of all three

[3] The MeToo hashtag gained traction on social media in 2017 and was associated with a movement calling attention to sexual assault and harassment (Walsh 2020; Bogen et al. 2021).

communities of practice also touched on the idea of purchasing or using commercially available data as a strategy for resolving intellectual property 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).

6.2.6.1 Intellectual Property and Data Ownership—Qualitative Researchers

Several qualitative researchers discussed data sovereignty and ownership when considering sharing or reusing qualitative data (qr = 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 (qr = 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 (qr = 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 CC-BY [licensed with a 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 [intellectual property] remains with the [data creators]" (QR01). However, another qualitative researcher believed that data were 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 intellectual property rights, these conflicting quotes from participants illustrate the participants' limited understanding of intellectual property and data ownership, especially regarding how they apply to data sharing and reuse.

6.2.6.2 Intellectual Property and Data Ownership—Big Social Researchers

Big social researchers were most concerned with intellectual property 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 (bsr = 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 (bsr = 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 intellectual property and data ownership 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).

6.2.6.3 Intellectual Property and Data Ownership—Data Curators

Data curators had similar concerns and strategies to other participants when dealing with issues of intellectual property and data ownership. They discussed social media terms of service—both following them and bending them (dc = 5). Data curators also talked about purchasing or using commercially available data (dc = 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 [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 intellectual property 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 (dc = 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 (dc = 2), and data licensing (dc = 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 intellectual property and data ownership (dc = 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).

6.2.7 Domain Differences

The term "domain" is a term used by Wenger et al. (2002) to describe the combination of interests and disciplines that are present within a community of practice. During the interview process, qualitative researchers, big social researchers, and data curators all discussed unique behaviors, attitudes, and practices that developed within their communities of practice due to the unique interests and disciplines that were shared within each community. This theme of domain differences therefore emerged during my coding process. 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. Qualitative researchers talked about collaborating with big social researchers, and vice versa, to support scaled-up, responsible research (n = 4; qr = 1, bsr = 3, dc = 0).

6.2.7.1 Domain Differences—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 (qr = 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 element of the analysis, saying "a core part of qualitative research is the idea of researcher as instrument" (QR03).

Qualitative researchers described a general reticence among their community of practice regarding sharing data. However, many of the qualitative researchers I spoke with were interested in the idea of sharing (qr = 7). One qualitative 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 [data], as long as they're going to be responsible and respectful?' I feel like that's making more use of [the data]" (QR03).

Only one qualitative researcher whom I spoke with discussed collaboration with big social researchers (qr = 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).

6.2.7.2 Domain Differences—Big Social Researchers

The big social researchers I interviewed discussed different practices and standards of different communities of practice (bsr = 4). For instance, some researchers described a

potential conflict between the common practice in their community and their own sense of responsibility, but they ultimately chose to stay aligned with other researchers in their area. As one researcher of recommender systems 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 into natural disasters:

> 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)

Big social researchers were interested in the idea of collaborating with social scientists to support responsible practice (bsr = 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 (bsr = 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 (bsr = 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)

6.2.7.3 Domain Differences—Data Curators

Data curators were able to speak about the differences between qualitative researchers and big social researchers from an outside perspective. Among the ten 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 (dc = 4), one data curator suggested that "big data [researchers] 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 (dc = 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).

6.2.8 Strategies for Responsible Practice

Another theme that emerged during my 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). Specific practices discussed by each participant group are explored below.

6.2.8.1 Strategies for Responsible Practice—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, the qualitative researchers I spoke with 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 (qr = 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 [at an early stage of my PhD program] 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).

6.2.8.2 Strategies for Responsible Practice—Big Social Researchers

Like qualitative researchers, big social researchers weighed a variety of risks and benefits as they conducted their research (bsr = 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 (bsr = 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 (bsr = 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 (bsr = 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).

6.2.8.3 Strategies for Responsible Practice—Data Curators

Like the other communities of practice in this study, data curators also discussed risk–benefit analysis. Data curators especially focused on 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, [and how the risk of harm affects the] recommendations we would make to deidentify the data further" (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)

Data curators also discussed talking with colleagues to help them make difficult curation decisions. As one curator said, "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).

6.2.9 Perspectives on Data Curation and Sharing

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 how participants perceived 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 areas of overlap, however, the three communities of practice had different ideas and concerns regarding data curation. This may indicate that communication between communities of practice would support stronger curation 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 indicates an ability for data curators to begin to bridge the gap between these two communities of practice.

6.2.9.1 Perspectives on Data Curation and Sharing—Qualitative Researchers

Several qualitative researchers emphasized the value of qualitative data sharing (qr = 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 enhance 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 (qr = 4) in order to support responsible data sharing. One researcher described how a consultation with a data librarian made them feel more comfortable 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 [the librarian's] help I was like, okay" (QR03). Another researcher described the benefits of their institutions' Qualitative Data Repository (QDR) membership: "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 (qr = 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 (qr = 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). However, some qualitative researchers (qr = 2) told me that the most important goal of qualitative data sharing is transparency, not data 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).

6.2.9.2 Perspectives on Data Curation and Sharing—Big Social Researchers

Big social researchers discussed the value of big social research (bsr = 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 a thousand 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 the community are being stressed, [and what resources] are being utilized during a crisis" (BSR06).

Big social researchers were also concerned with how the technical requirements of big social data can hinder sharing and reuse (bsr = 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.

6.2.9.3 Perspectives on Data Curation and Sharing—Data Curators

Data curators were strong believers in the value of data sharing (dc = 3). One data curator emphasized increased citations as an incentive to publish data: "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 (dc = 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 (dc = 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 in which researchers did not plan ahead for data curation, or cases in which data presented special 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 (dc = 2), as long as certain standards are met. As one curator said, "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 believed that sharing some amount of data for transparency purposes was better than sharing nothing (dc = 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).

6.3 Summary

The qualitative researchers, big social researchers, and data curators I spoke with were deeply engaged with the key issues held in common between 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. The interviews highlighted both similarities and differences in how each community approached these issues. Additionally, insights arise from the three new themes I identified in my content analysis—domain differences, strategies for responsible practice, and perspectives on data curation and data sharing. Particularly, the interviews showed a disconnect between the three communities of practice. Even with the rise interdisciplinarity as a trend, domain and disciplinary assumptions and silos were present among the participants I interviewed. It was rare for qualitative researchers and big social researchers to interact, and therefore any strategies for responsible practice were developed extemporaneously, in consultation with others within their discipline or domain. While data curators have the potential to bolster responsible practice by connecting researchers and providing services, each community of practice had different interests, concerns, and assumptions regarding data curation and data sharing. I will explore these ideas further in Chap. 7.

References

Bogen KW, Bleiweiss KK, Leach NR, Orchowski LM (2021) #MeToo: disclosure and response to sexual victimization on Twitter. J Interpers Violence 36:8257–8288. https://doi.org/10.1177/0886260519851211

Chang Y-W (2018) Exploring the interdisciplinary characteristics of library and information science (LIS) from the perspective of interdisciplinary LIS authors. Libr Inf Sci Res 40:125–134. https://doi.org/10.1016/j.lisr.2018.06.004

Mannheimer S (2023) Interviews regarding data curation for qualitative data reuse and big social research. Qualit Data Repository.https://doi.org/10.5064/F6GWMU4O

Perma.cc (2023) About Perma.cc. https://web.archive.org/web/20230321151818/http://perma.cc/about

Walsh C (2020) Challenge of archiving the #MeToo movement. Harvard Gazette

Insights from Interviews with Researchers and Curators

7

In this chapter, I discuss insights drawn from my interviews with qualitative researchers, big social researchers, and data curators, focusing on similarities and differences between communities of practice, and discussing implications for data curation. The initial discussion is organized around the six key issues that have structured this book—context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. I then discuss new ideas that emerged from the interviews about domain differences, strategies for responsible practice, and perceptions on data curation and sharing. This chapter concludes with a discussion of implications for data curation practice.

7.1 Original Six Issues Drawn from the Existing Literature

7.1.1 Context

In the interviews, I asked researchers and curators 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. However, there were also key differences in how each community of practice considered how to preserve contextual information.

S. Mannheimer, *Scaling Up: How Data Curation Can Help Address Key Issues in Qualitative Data Reuse and Big Social Research*, Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-49222-8_7

Qualitative researchers were concerned with how to communicate the deep context inherent in qualitative research—for example, how research is co-created with participants, how a researcher's background affects context, and the details of the community where the study took place. 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 concerned with the time required to fully document context, and they were concerned that providing context-enhancing details could potentially endanger participant privacy and confidentiality. Big social researchers, on the other hand, were more focused on the more technical aspects of context—for example, 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, big social researchers tended to view contextual issues with less concern than qualitative researchers. Big social researchers acknowledged these issues as a part of their research, but none whom I interviewed thought these issues would compromise their research.

Data curators focused on how context can be enhanced by clear documentation, rich description, standardized metadata, and links to related materials. Data curators were also able to speak to the similarities and differences between qualitative and big social data. For example, they emphasized that qualitative data required more in-depth review and description than big social data, and they were also concerned about the potential participant privacy implications of providing too much contextual information for both qualitative data reuse and big social research.

7.1.2 Data Quality and Trustworthiness

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 of practice also discussed data completeness as an important element of quality and trustworthiness, especially the importance of communicating the level of data completeness or missing data. However, each community of practices also had unique considerations regarding data quality and trustworthiness that were wide-ranging and specific to the type of data being analyzed or collected.

Qualitative researchers were concerned with the human aspects of data quality—they described how they documented data quality issues in manuscripts, they were concerned with researcher bias, they considered the trustworthiness of data creators, and they noted that nuances of human communication can be 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 sharing code and related documentation to support quality and trustworthiness. While all three communities of practice were concerned with fully describing data quality issues to support research integrity and data reuse, data curators discussed this the most. 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 the interviews, I asked the participants to describe challenges relating to comparing and combining different datasets. My review of existing literature suggests that comparing and combining data can enable higher quality research (e.g., larger scale of research, more representative samples, broader conclusions). And indeed, all three communities of practice discussed how comparing and combining data can yield stronger research and conclusions. However, combining datasets is made more difficult for qualitative researchers and big social researchers because of challenges relating to missing data, research questions, methods, and metadata interoperability.

Qualitative researchers, big social researchers, and data curators all understood the theoretical value of comparing and combining datasets to support broader conclusions and more representative samples. However, in practice, many of my interviewees with were thwarted by challenges that prevent comparability—for example, different data formats and different metadata formats. A few big social researchers I spoke with had successfully combined datasets, especially to support demographic information and more representative study populations. However, no qualitative researchers I spoke with had done so. Because each community of practice had different levels of experience and different concerns and focuses regarding data comparability, this appears to be an area in which connecting communities of practice could be most beneficial. Big social researchers' experience with this practice, along with data curators' expertise in metadata and format interoperability, could be applied to support qualitative researchers who wish to compare and combine qualitative datasets.

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 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 usually 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. However, other than topics relating to IRBs, the concerns of qualitative researchers and big social researchers regarding informed consent did not overlap. This 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.

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 used strategies such as broad consent, tiered consent, and restricted access to mitigate potential consent issues stemming from data access and reuse. 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.

Conversely, while a few big social researchers had considered the problematic nature of consent for big social data, others told me that they did not consider their research to be human subjects research at all, and therefore informed consent was unnecessary. Regardless of their perspective on consent, 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. Big social researchers generally considered these 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.

The data curators I spoke with were conversant in the issues that mattered to both qualitative data reuse and big social research, suggesting that data curators are well-positioned to build connections between communities of practice. Data curators described using several different strategies to protect participants even if informed consent was not obtained—for example, ensuring deidentification of data, providing restricted access, considering the sensitivity of data, and providing data enclaves where reusers can analyze data without downloading it. Data curators also discussed the importance of connecting with researchers early in the research process as the key strategy for supporting consent. At this early stage, with the right training (see Chap. 8, Sect. 8.2.2), 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. 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, including consideration of informed consent.

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. The three communities of practice were fairly consistent in how they understood and addressed the issue of privacy and confidentiality.

Qualitative researchers were fluent in issues of privacy and had implemented various strategies for preserving the privacy and confidentiality of their research participants. Big social researchers were also highly concerned about participant privacy and confidentiality; in fact, they viewed privacy protection as even more important because they did not generally obtain informed consent from participants. When considering privacy and confidentiality, all three communities of practice discussed data deidentification, data sensitivity, restricted access, participant/user expectations of privacy, potential harms to participants, research design for privacy, and data security. This finding suggests that privacy-focused data curation strategies are applicable to both qualitative data and big social data.

7.1.6 Intellectual Property and Data Ownership

In this study, I asked participants to describe challenges relating to intellectual property and data ownership. Participants generally had limited understandings of intellectual property and data ownership, and few had considered these issues in detail.

Most qualitative researchers had not considered the intellectual property rights or data ownership of research participants, and these concerns did not greatly affect their practices of data sharing and reuse. On the other hand, most big social researchers were aware of the impact of platform terms of service when collecting big social data. While a few of the big social researchers I spoke to described purposefully breaking terms of service, most felt obligated to adhere to any big social data terms of service. In complying with such terms of service, the majority of big social researchers I spoke to had not shared their research data publicly, opting instead to describe their data collection methods so that future researchers could replicate the data collection process for themselves.

Data curators were the most fluent in intellectual property rights and data ownership concerns for both qualitative and big social data. Many data curators I spoke with discussed data licensing, data citation, and curatorial review for intellectual property and data ownership concerns. Some had also helped researchers find research data for reuse and had facilitated purchasing commercially available data. The data curators I spoke with also discussed addressing 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.

7.2 Additional Themes

As I wrote in Chap. 6, three additional themes emerged from the interviews. First, domain differences—that is, differences in how each community of practice considered each of the interview prompts, based on the interests, disciplines, values, and research methodologies within that community of practice. In my discussion of domain differences, I also explore how each community of practice had different focuses and approaches to each issue, and I discuss how different communities of practice had different viewpoints about whether reused data should be viewed as human subjects data or as unembodied "content." Second, I discuss the strategies that interview participants have developed for ethical, legal, and epistemologically sound research (referred to in shorthand as "responsible research"). Third, I discuss the each community of practice's perspectives on data curation and sharing.

7.2.1 Domain Differences

Within each community of practice, there was generally alignment regarding approaches and prioritization of key issues, due to the similar domains of the members of each community of practice—that is, the intersection of their disciplines, interests, values, and research methodologies. However, the domain differences between the communities of practice led to different approaches, values, and viewpoints, and different skills and training. 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 domains to support scaled-up, responsible research. However, few participants reported specific instances of connecting with researchers from other communities of practice—and for those who did, the researchers from other communities of practice served in consultant roles, not as full collaborators.

In this research, domain differences manifested in two key ways: different focuses and approaches to issues, and different viewpoints on what constitutes human subjects data. I discuss each of these ideas below.

7.2.1.1 Different Focuses and Approaches to Each Issue

While the interviews showed that the six key issues identified in Chaps. 3 and 4 (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 viewed each issue through a domain-specific lens, and therefore had different focuses and approaches for each issue.

Using the issue of context as an example: As noted above in Sect. 7.1.1., qualitative researchers were trained to consider how their data analysis might be affected by the complexities of participants' (and researchers') life experiences and perspectives. On the other hand, big social researchers were accustomed to the idea that 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.

These different focuses and approaches demonstrate the value of connecting the three communities of practice. 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, should 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, in Sect. 7.3.1.

7.2.1.2 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, were more likely to think of big social data as unembodied "content," rather than as an extension of the human participants who created that content. This foundational philosophical mismatch between qualitative researchers and big social researchers provides insight into key differences between the two communities' approaches to research. The issue of consent provides an illustrative example. As noted above in Sect. 7.1.4, qualitative researchers were concerned about participant consent for research with archived or reused data, considering 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.

However, when considering the issue of privacy and confidentiality, both big social researchers and qualitative researchers were aligned, and all three communities of practice used similar data curation strategies to ensure privacy (see Sect. 7.1.5, above). So even

as big social researchers may consider big social data to be unembodied content as they collect those data, they also recognize the importance of protecting the privacy of people represented in their research data. This alignment on the issue of privacy may be an opportunity for data curators to engage with big social researchers. Data curators can connect with big social researchers to help them preserve privacy and confidentiality in their research. During those interactions, data curators can also check in with big social researchers on the other issues discussed in this book.

7.2.2 Strategies for Responsible Practice

The participants I interviewed often drew upon many sources to cobble together strategies for responsible practice. Qualitative researchers, big social researchers, and data curators all described a process of continuous re-examination of epistemological, ethical, and legal issues—making decisions on the fly about how to act responsibly. Researchers and data curators 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 discussed IRBs as potential partners for ethical concerns. However, when research uses existing data (including qualitative data reuse and big social research), IRBs generally either do not require review or grant exempt status. It was rare for participants to discuss any other 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 the big social research community, not the qualitative data sharing and reuse community. The big social researchers I interviewed came from a variety of disciplinary backgrounds (civil engineering, communication, computer science, information science, journalism, and 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.

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 Perspectives on Data Curation and Data Sharing

During their interviews, participants often discussed the broad benefits and significant challenges of data curation. While several 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 data curation considerations can help data curators better serve these communities of practice by providing tailored data curation resources that respond directly to researchers' needs. Understanding researchers' different needs and priorities can also enable data curators to better advocate for data sharing, despite the time and effort required. Data curation is an area in which communication between communities of practice could support stronger practices, and data curators are well-positioned to act as a bridge between qualitative researchers and big social researchers.

7.3 Implications for Data Curation Practice

Through my exploration of the similarities and differences between how key issues are discussed by big social researchers, qualitative researchers, and data curators, new data curation insights emerge. For many of the issues discussed in this book, data curation can help enhance responsible practice. In talking with members of each community of practice, it also became clear that data curators can act as facilitators and intermediaries to connect qualitative researchers with big social researchers to encourage responsibly scaling up social research.

Data curators were able to speak fluently about a variety of issues—both issues that concerned big social researchers and issues 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 intellectual property law meant and how it impacted their research, but they were aware of how data curation could support intellectual property 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. Of course, these data curation services require that data curators have the appropriate expertise. I discuss the importance of training for qualitative researchers, big social researchers, and data curators in the next chapter, in Sect. 8.2.2. Table 7.1 provides an overview of the six key issues, the aspects of each issue addressed by data curators in their interviews, and the applicable data curation strategies that curators can use to address each issue.

7.3.1 Planning Ahead for Data Curation

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 data repositories are available to support responsible research practices. 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. Data curators confirmed this from their end, telling me that it is difficult to reach researchers early in the research process.

The importance of planning ahead for data sharing is widely acknowledged in the scientific community, as notably illustrated by U.S. federal funders' requirements of data management and data sharing plans in grant proposals. However, when researchers write data management plans for grant proposals, they don't always consult with data curators or data repositories, and even if they do have contact with data curators during the grant

Table 7.1 Aspects of issues addressed by data curators and coinciding data curation strategies

Issue	Data curator focuses	Data curation strategies
Context	Documentation and related materials	• Work with researchers to include in-depth documentation, metadata, and linked materials alongside datasets in repositories
Data quality and trustworthiness	Repository trustworthiness, and quality of metadata and documentation	• Work with researchers to create thorough, high-quality metadata and documentation • Pursue certifications for trustworthy repositories and/or align with TRUST Principles • Check data and code to ensure it is readable and executable
Data comparability	Metadata and format interoperability	• 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	Responsibility of data repositories, providing access to shared data whenever as possible	• 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

(continued)

Table 7.1 (continued)

Issue	Data curator focuses	Data curation strategies
Privacy and confidentiality	Repository and curator support for privacy	• 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	Intellectual property as it relates to data repositories	• Training for researchers on intellectual property concepts • Repository terms of use • Data citation • Data licensing • Guidance on data sovereignty, ownership, and governance • Rights clearance and management for reused datasets

proposal process, they may not re-engage with data curators at the outset of a funded grant.

Beyond data management plans, my research suggests a few strategies for early contact between data curators and researchers. 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. Going even further, data curators could potentially work with these partners to implement data curation requirements—for example, IRBs could require consultation with a data curator prior to granting exempt status to big social research or qualitative data reuse projects, or university research support offices could require a consultation with a data curator prior to dispensing grant funds.

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 promote 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.

7.4 Chapter Summary

This research shows that qualitative researchers and big social researchers, as distinct communities of practice, are under-connected. While some participants told me 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 there is an opportunity for data curators to build connections between these two other 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. The issues identified in this research 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. This broad knowledge also position data curators 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 Sect. 7.3. However, data curators as a community of practice are also under-connected with qualitative researchers and big social researchers. This under-connection means that the qualitative researchers and big social researchers I spoke with relied on informal strategies to support responsible practice, rather than reaching out to data curators for help. Encouraging connection between all three of these communities of practice and planning ahead for research and data sharing will support more responsible research and enhanced data sharing, thus leading to additional discoveries and insights in behavioral and social science.

Scaling Up Responsibly: Connecting Communities of Practice

8

In this book, I have explored the connections between qualitative data reuse, big social research, and data curation. I reviewed existing literature to identify the key issues of context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. Then I interviewed qualitative researchers, big social researchers, and data curators to further dive into each key issue and arrive at new insights about how domain differences affect each community of practice's viewpoints, different strategies that researchers and curators use to ensure responsible practice, and different perspectives on data curation. This chapter outlines key contributions of the research, ideas for future work, and closing thoughts about scaling up qualitative research.

8.1 Contributions

This research ultimately shows that the three communities of practice I investigate—qualitative researchers, big social researchers, and data curators—are under-connected. While all three communities of practice are affected by many of the same key issues when conducting big social research, qualitative data reuse, and data curation, these three communities of practice often emphasize different aspects of those issues. The fact that the members of each community of practice often addressed different aspects of each issue is precisely why it would be beneficial for these three communities to come together to help responsibly scale up their research. The different aspects of these issues are key to connecting research communities and data curators for their mutual benefit: the focuses and

S. Mannheimer, *Scaling Up: How Data Curation Can Help Address Key Issues in Qualitative Data Reuse and Big Social Research*, Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-49222-8_8

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, forming research questions that produce important insights without posing undue risk to participants, carefully deidentifying direct quotes, and using 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 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 as a method for scaling up qualitative research and building 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 and legal 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, data licensing, 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 interviewed for this book rarely contacted data curators before their research was complete and they were actively considering sharing their data. And many of the big social researchers I spoke with never contacted data curators at all. This meant that the researchers were not able to benefit from data curators' knowledge during the research process; instead, they cobbled together informal strategies to support responsible practice. This research 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 research also highlights areas in which data curators can offer specific services—for example, data curators can provide guidance on consent procedures that support data reuse, help navigate social media terms of service, or assist with big social research design. Data curators may need additional training to provide these services. See Sect. 8.2.2 for further discussion on education and skills-building.

8.2 Future Work

The research presented in this book could lead to future work in several different areas—both research-focused and practice-focused. I discuss a few ideas below.

8.2.1 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. Advocating for new legislation and regulation may help, but in the meantime, the scholarly community needs to find ways to ensure epistemologically

sound, ethical, and legal big social research and qualitative data reuse. The fact that only two researchers whom I interviewed 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 (RDAP), Digital Library Federation, and International Association for Social Science Information Service & Technology (IASSIST) 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 (discussed further below, in Sect. 8.2.2.). As with any standard, the community will need to commit to regularly revising and updating these standard practices.

8.2.2 Education and Skills Development

Most of the researchers interviewed for this book reported that they had not received specific training and guidance about the epistemological, ethical, and legal issues inherent in qualitative data reuse and big social research practices. As these types of research become more commonplace, graduate programs have begun to adapt their curricula to address data reuse, emerging data types, artificial intelligence, and big data analytics (e.g., Clayton and Clopton 2019; Haaker 2020; Giacomello and Preka 2020; Bond 2022). It has also become more common for universities to offer graduate certificates in data science and data analytics, aimed at researchers and professionals who want to enhance their skills in this area (Jiang and Chen 2022). However, such graduate programs generally offer standalone courses on technology ethics, rather than infusing ethics and responsible research practices into the entire curriculum (Grosz et al. 2019; Fiesler et al. 2020). We need more thoughtful curriculum development that integrates the skills and competencies necessary for truly responsible research practice.

While the data curators I spoke with suggested a variety of strategies for dealing with key issues, knowledge of these strategies is not widespread among the data curation community. Data curators will need additional training if they are to tailor their services to address the key issues identified in this book. Data curators are usually trained in Library and Information Science graduate programs, where curriculum focuses on data organization, metadata, digital preservation, and access considerations. Data curators need more specialized skills if they are to provide services such as reviewing informed consent procedures or interpreting intellectual property rights. These skills are beginning to be developed in the data curation community. Ithaka S + R recently convened cohorts

of librarians, researchers, and IT professionals to investigate big data research needs and services at universities (Lutz 2021). Library and Information Science programs are beginning to include courses that address services to support data science, big data research, and data reuse (Urs and Minhaj 2022). Data curators who specialize in qualitative data and big social data are beginning to share their expertise via documentation and publications (e.g., Hemphill et al. 2018; Demgenski et al. 2021). However, these initiatives are still in the early stages, and more training is necessary, both during graduate school and in the form of professional development for working data curators.

8.2.3 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 book; this issue warrants additional research to develop strategies for preserving context in both qualitative and big social data. And more research and advocacy could be done to develop and operationalize interoperable, standardized metadata schemas, thus enhancing data comparability.

8.2.4 The Changing Big Social Research Landscape

Big social research methods and big social data sources are 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. The popularity of different social media and online spaces are constantly evolving: new types of platforms are emerging, and social media influencers have become a prominent user group in recent years. Elon Musk's acquisition of Twitter in 2022 (Chotiner 2022; Conger 2023) 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. Big social research methods are also continually evolving. New technologies, tools, and systems such as new artificial intelligence models and ever-more-powerful supercomputers expand the complexity and capacity of big social research. Researchers and data curators will need to contend with this rapidly-evolving landscape, which will affect all six key issues addressed

in this book—context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. Future research could investigate how key issues change with the changing online spaces and research technologies. Data curation strategies may also need to be adjusted to support evolving data sources and technology innovations in big social research.

8.2.5 The Value of Small Data

This book operates under the assumption that scaling up research is an important goal. As Kitchin (2014) 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." 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." However, Kitchin (2014) 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." 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 research answers different questions from ethnographic research or other types of in-depth social research. Housley et al. (2014) 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."

Scaling up research may not always be the goal. As Boyd and Crawford (2012) write, "The size of data should fit the research question being asked; in some cases, small is best"—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 book. 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.

8.3 Closing Thoughts

As data sharing continues to grow and social research continues to scale up, the key issues discussed in this book will evolve in scope and complexity. Throughout the book, I have focused on the role of data curation in supporting responsible research and data sharing. Data curation is a growing profession, and an increasing number of trained curators are well-positioned to lead data curation initiatives. Data curation practices can also 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. 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 can begin to bridge gaps between these communities. Encouraging these connections between different communities of practice can help us increase data sharing and reuse and responsibly scale up social research, ultimately enhancing discoveries in social and behavioral science.

References

Bond R (2022) Building a foundation for data science researchers in political science. In: Brown M, Nordyke S, Thies C (eds) Teaching graduate political methodology. Edward Elgar Publishing, Cheltenham, pp 212–217

Boyd D, Crawford K (2012) Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inf Commun Soc 15:662–679. https://doi.org/10.1080/1369118X.2012.678878

Chotiner I (2022) Why Elon Musk bought Twitter. The New Yorker

Clayton PR, Clopton J (2019) Business curriculum redesign: Integrating data analytics. J Educ Bus 94:57–63. https://doi.org/10.1080/08832323.2018.1502142

Conger K (2023) How Elon Musk is changing the Twitter experience. The New York Times

Demgenski R, Karcher S, Kirilova D, Weber N (2021) Introducing the qualitative data repository's curation handbook. J eSci Librariansh 10:1207. https://doi.org/10.7191/jeslib.2021.1207

Fiesler C, Garrett N, Beard N (2020) What Do We Teach When We Teach Tech Ethics?: a Syllabi Analysis. In: Proceedings of the 51st ACM technical symposium on computer science education. ACM, Portland, OR, pp 289–295

Giacomello G, Preka O (2020) The "social" side of big data: teaching BD analytics to political science students. Big Data Cogn Comput 4:13. https://doi.org/10.3390/bdcc4020013

Grosz BJ, Grant DG, Vredenburgh K, Behrends J, Hu L, Simmons A, Waldo J (2019) Embedded EthiCS: integrating ethics across CS education. Commun ACM 62:54–61. https://doi.org/10.1145/3330794

Haaker M (2020) Qualitative secondary analysis in teaching. In: Hughes K, Tarrant A (eds) Qualitative secondary analysis. Sage Publications, Los Angeles, CA, pp 119–134

Hemphill L, Leonard SH, Hedstrom M (2018) Developing a social media archive at ICPSR. In: Web archiving and digital libraries (WADL). Fort Worth, TX

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 Soc 1. https://doi.org/10.1177/2053951714545135

Jiang H, Chen C (2022) Data science skills and graduate certificates: a quantitative text analysis. J Comput Inf Syst 62:463–479. https://doi.org/10.1080/08874417.2020.1852628

Kitchin R (2014) The data revolution: big data, open data, data infrastructures and their consequences. Sage Publications, Los Angeles, CA

Lutz K (2021) Supporting big data research. In: Ithaka S+R. https://sr.ithaka.org/blog/supporting-big-data-research/

Manovich L (2012) Trending: the promises and the challenges of big social data. In: Gold MK (ed) Debates in the digital humanities. University of Minnesota Press, Minneapolis, MN, pp 460–475

Urs SR, Minhaj M (2022) Evolution of data science and its education in iSchools: an impressionistic study using curriculum analysis. J Assoc Inf Sci Technol asi.24649. https://doi.org/10.1002/asi.24649