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

Introduction (1) -- Insights from Interviews with Researchers and Curators (Ch 7).

Sara Mannheimer

Mannheimer, S. (2024). 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: Springer.https://doi.org/10.1007/978-3-031-49222-8

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

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.