Data Narrative

Interviews regarding data curation for qualitative data reuse and big social research

Sara Mannheimer

Project Summary

Trends toward open science practices, along with advances in technology, have promoted increased data archiving in recent years, thus bringing new attention to the reuse of archived qualitative data. Qualitative data reuse can increase efficiency and reduce the burden on research subjects, since new studies can be conducted without collecting new data. Qualitative data reuse also supports larger-scale, longitudinal research by combining datasets to analyze more participants. At the same time, qualitative research data can increasingly be collected from online sources. Social scientists can access and analyze personal narratives and social interactions through social media such as blogs, vlogs, online forums, and posts and interactions from social networking sites like Facebook and Twitter. These big social data have been celebrated as an unprecedented source of data analytics, able to produce insights about human behavior on a massive scale. However, both types of research also present key epistemological, ethical, and legal issues. This study explores the issues of context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership, with a focus on data curation strategies. The research suggests that connecting qualitative researchers, big social researchers, and curators can enhance responsible practices for qualitative data reuse and big social research.

This study addressed the following research questions:

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

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

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

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

Data Description and Collection Overview

The data in this study was collected using semi-structured interviews that centered around specific *incidents* of qualitative data archiving or reuse, big social research, or data curation. The participants for the interviews were therefore drawn from three categories: researchers who have used big social data, qualitative researchers who have published or reused qualitative data, and data curators who have worked with one or both types of data. Six key issues were identified in a literature review, and were then used to structure three interview guides for the semi-structured interviews. The six issues are context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership.

This semi-structured interview process was informed by the guidelines for information science researchers laid out by Luo and Wildemuth (2017). To conduct the interviews, three interview guides—one for qualitative researchers, one for big social researchers, and one for data curators were created, including “essential questions, extra questions, throw-away questions, and probing questions” (*ibid*., p. 234).

Interview participants were selected differently based on participant type. Participants who were data curators were identified through the study’s literature review by contacting authors of key articles. Additional data curators were contacted as a result of the author’s knowledge of the data curation community. Emphasis was placed on those who had experience with qualitative and big social data.

Two strategies were used to select participants who were qualitative researchers. First, the author searched the Qualitative Data Repository (Center for Qualitative and Multi-Method Inquiry, 2020) for datasets that had been published in the last four years to identify participants who had published qualitative data. Similarly, the author also searched Dryad (Dryad, 2022) and Zenodo (CERN Data Centre, 2020) for qualitative data, using the keywords “interview” and “qualitative.” For the second strategy, the author searched the Web of Science database for the keywords “qualitative data reuse” and “qualitative secondary analysis,” then filtered for articles published in the past four years to identify researchers who had recently reused qualitative data. The final group of qualitative researchers included two researchers who had reused qualitative data from other sources, three researchers who had conducted secondary analysis on their own qualitative data, and five researchers who had shared their own qualitative data in a repository.

For big social data researchers, the author searched Web of Science for the keywords “big social data,” “social media data,” “social media,” “facebook,” “twitter,” “reddit,” and “pinterest”, then filtered for articles published in the past four years. The author was able to identify additional interviewees by asking dissertation advisors and mentors for suggestions.

Additional participants were found using snowball and theoretical sampling. There was a higher initial response rate from qualitative researchers who had published data in a repository, and the author noted a gap in the analysis regarding the viewpoints of participants who had reused qualitative data; the author therefore purposefully searched Web of Science for additional participants who had reused qualitative data. Sampling continued until saturation was reached—that is, “the point in the research when all the concepts are well defined and explained” (Corbin & Strauss, 2008, p. 145).

Participants were limited to those working in the United States. Ten participants from each of the three target populations—big social researchers, qualitative researchers who had published or reused data, and data curators were interviewed. The interviews were conducted between March 11 and October 6, 2021. The longest interview lasted one hour and 13 minutes and the shortest lasted 33 minutes, with an average interview length of 53 minutes.

All interviews were scheduled using Microsoft Bookings software (Ako-Adjei & Penna, 2021) and conducted using Zoom videoconferencing software (Zoom, 2021). When scheduling the interviews, participants received an email asking them to identify a critical incident prior to the interview. This strategy is derived from critical incident technique, in which an interviewer uses a specific example, or "incident," to focuses a participant’s answers to the interview questions. This focus allows participants to remember more detail and provide concrete examples and experiences (Flanagan, 1954). The email also provided them with the IRB-approved consent agreement for their review and the full text of the applicable interview guide, which included a short description of the research.

The author began each interview with an introduction and gave an overview of the research being conducted—reading from the description that was provided on the first page of the interview guide. It was then explained to the participant that there would be eight question areas—an introductory section, one section for each of the six key issues identified in the literature review, and a wrap-up section. The participants were asked their permission to have the interviews recorded, which was completed using the built-in recording technology of Zoom videoconferencing software (Zoom, 2021). The author also took notes during the interviews.

Otter.ai speech-to-text software (Otter.ai, 2021) was used to create initial transcriptions of the interview recordings. A hired undergraduate student hand-edited the transcripts for accuracy. The student made notes when they had questions or when the recording was unclear, and the author conducted a final review of the transcripts for accuracy. The author also conducted an initial deidentification of the transcripts at this stage, in the summer and fall of 2021. Additional deidentification was conducted in partnership with the curators at the Qualitative Data Repository, where the transcripts are shared.

The author analyzed the interview transcripts using a qualitative content analysis approach. This involved using a combination of inductive and deductive coding approaches, as outlined in Zhang and Wildemuth (2017) and as detailed in Bernard, Wutich, and Ryan (2017). After reviewing the research questions, the author used NVivo software to identify chunks of text in the interview transcripts that represented key themes of the research (QSR International, 2022). Because the interviews were structured around each of the six key issues that had been identified in the literature review, the author deductively created a parent code for each of the six key issues. These parent codes were context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. The author then used inductive coding to create subcodes beneath each of the parent codes for these key issues.

Selection and Organization of Shared Data

The data files consist of the interview transcripts themselves – transcripts from Big Science Researchers (BSR), Data Curators (DC), and Qualitative Researchers (QR) respectively. Additional data files include the redacted interview analysis, and participant summaries (Memos). The documentation files include the individual interview guides for each of the participant categories, two versions of the consent form, a codebook, the emails sent to the participants, a file of the interview dates and duration, the IRB exempt application.

REFERENCES:

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

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

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

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

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

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

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*(4), 327–358. https://doi.org/10.1037/h0061470

Luo, L., & Wildemuth, B. M. (2017). Semistructured interviews. In *Applications of Social Research Methods to Questions in Information and Library Science* (2nd ed., pp. 248–257). Libraries Unlimited.

Otter.ai. (2021). *Otter.ai speech-to-text software*. https://web.archive.org/web/20220101184238/https://otter.ai/

QSR International. (2022). *NVivo qualitative data analysis software*. https://web.archive.org/web/20220402173708/https://www.qsrinternational.com/n vivoqualitative-data-analysis-software/home

Zhang, Y., & Wildemuth, B. M. (2017). Qualitative analysis of content. In *Applications of Social Research Methods to Questions in Information and Library Science* (2nd ed.). Libraries Unlimited.

Zoom. (2021). *Video conferencing, cloud phone, webinars, chat, virtual events*. https://web.archive.org/web/20220102000820/https://zoom.us