Protecting Participant Privacy While Maintaining Content and Context: Challenges in Qualitative Data De-identification and Sharing

Claire A. Myers, Shelby E. Long, and Faye O. Polasek, University of Michigan, USA
clairemy@umich.edu, longshe@umich.edu, fpolasek@umich.edu

Abstract

The Library Assessment for Research and Scholarship Lab investigates qualitative research support across disciplines. In 2018-2019, the lab conducted 29 interviews with faculty, librarians, and doctoral students who engaged in qualitative research to understand their needs during the research lifecycle. At the conclusion of this project, the qualitative data will be deposited in a repository where it can be made available for future secondary use. The deposited data will include de-identified versions of the complete interview transcripts. This poster supplements existing de-identification standards, details drafting and revising protocol for de-identification of our data, and discusses the de-identification process we used for the qualitative data. Existing de-identification literature and standards are limited and not widely uniform in qualitative research. In developing de-identification protocol, our lab recognized several potential challenges in the process and created procedures to ensure future data usability. There is inherent tension between keeping privacy intact and sharing undistorted qualitative data. We aim to address some of the hazards with de-identification best practices, demonstrating methodology for producing high quality de-identified qualitative data. In offering up a test case with suggested methods to better protect participants' identities, this work will lend itself to sustainable qualitative data sharing and reuse.

Background

Our lab conducted interviews to inquire about the data management and data sharing practices of researchers using qualitative and mixed methods. In the process we were generating qualitative data of our own. We decided to make our qualitative interviews available to future researchers, committing to the important work of careful de-identification for accessible data sharing. We hope to provide a starting point for future de-identification work by detailing our process and sharing our de-identification protocol.

Methods

We contacted experts from the Qualitative Data Repository (QDR) at Syracuse University early on in the data gathering phase of our research to discuss the task of preparing our data for deposit and gaining a basic understanding of the requirements. When we completed the data gathering phase, additional resources on qualitative data de-identification were gathered and discussed including CESSDA's "Data

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pra2.415

Management Expert Guide," University of Michigan's "Data Security Guidelines," and ICPSR's "Guide to Social Science Data Preparation and Archiving."

[Image 1]

Ultimately we decided to follow more closely the De-Identification guidelines provided by QDR. More specifically we adhered to their suggestions to keep a log of every alteration, develop a protocol, and document our process of creating and applying our protocol, which we outline in the summary below.

[Image 2]

We had to ensure that the de-identification protocol would be understood and applied consistently across the interview transcripts by all team members. We frequently had varying opinions on how to proceed with de-identification and what qualified as identifying information. In order to simplify our workflow, we decided to have one team member assigned to quality control. This person was responsible for reviewing each de-identified transcript and ensuring that the other team members were de-identifying and formatting the document consistently, allowing us to work at a faster pace.

[Image 3]

While QDR suggests best practices, they do not supply examples of de-identification logs and most guidelines also do not provide an example de-identification protocol. Recognizing that access to those resources and a guide for completing this process as a team would have been useful for us, we opted to attach our final protocol here to assist researchers in creating their own. It is important to note that researchers should anticipate an iterative process when developing their methods; however, our preliminary protocol offers an entry point into this necessary exercise.

[Images 4, 5, and 6]

Challenges

In addition to lacking specific examples of de-identification logs and team protocols, we were presented with specific de-identification challenges due to the nature of the data we were collecting. Participants included detailed information about the specific research questions that they were pursuing in their work. This was particularly difficult to work around, however, if we completed de-identification carefully the chances that individuals would be identified by their research questions decreased significantly.

[Image 7]

In one particular instance we left granular information about a researcher's current work mostly intact, but decided instead to remove their discipline and all other personal identifiers. We did so to maintain necessary context given how their specific academic focus shapes their research process. To test whether the participant's identity could be compromised, we searched online for their research focus and concluded we could leave it in when we were unable to identify the participant in our search. In other situations we may decide to retain the discipline and remove particularities instead. None of the guidelines we found suggested such a specific approach. This example demonstrates how the process varies depending on the data, even with solidified protocol.

[Image 8]

Discussion

In our de-identification work, we found that each decision involved weighing the potential risk to participant privacy with the desire to preserve the original context. This is consistent with the literature which describes the "trade-off between sharing and risk to privacy" (Kirilova & Karcher, 2017). Participants in our own research echoed these common apprehensions expressed by qualitative researchers across disciplines. However, we hope that by sharing our process for preparing our data for sharing, we demystify the procedure for other teams and can ease those concerns.

[Image 9]

The data we collected was not sensitive nor was it collected from a vulnerable population. The potential risk beyond breach of privacy was reputational harm as interviewees shared details from their professional experiences. The most effective de-identification processes address such harms directly, accounting for aspects of participant population information, such as vulnerable population status. Just as researchers consider these designations when drafting the consent forms, they must also inform the de-identification efforts.

While de-identification is detailed and time consuming work, it is a worthwhile endeavor which contributes to our engagement in data sharing and reuse.

[Image 10]

Conclusion

The Library Assessment for Research and Scholarship Lab will continue to de-identify qualitative interview transcripts with the goal of depositing the complete de-identified files in a qualitative data repository. Future researchers will not only be able to access the codebook and supplemental information describing the data, but the actual data itself thanks to our de-identification work.

The protocol that we developed can serve as a guide for other qualitative researchers who are interested in sharing their data. We used the guidelines provided by QDR, tailored them to our needs and the needs of our particular data set, and expanded on them. It is likely that other researchers will also have to make changes based on their unique data. With the work of QDR as a foundation and our complete protocol to provide additional guidance, we hope this work can ease the process for others and demonstrate that qualitative data de-identification is not only possible but can provide a path to qualitative data sharing and availability.

Acknowledgements

This research was made possible in part with funding from the Institute of Museum and Library Services grant #RE-95-17-0104-17

83rd Annual Meeting of the Association for Information Science & Technology | October 24-28, 2020.

References

CESSDA Training Team. (2020). CESSDA Data Management Expert Guide. Bergen, Norway: CESSDA ERIC.

Data Security Guidelines. (2020). Retrieved from https://research-compliance.umich.edu/data-security-guidelines

De-Identification. (n.d.). Retrieved from https://qdr.syr.edu/guidance/human-participants/deidentification

Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle 6th Edition. (n.d.). Retrieved from https://www.icpsr.umich.edu/files/deposit/dataprep.pdf

Kirilova, D. & Karcher, S. (2017). Rethinking data sharing and human participant protection in social science research: Applications from the Qualitative Realm. *Data Science Journal, 16*(43), 1-7. http://dx.doi.org/10.5334/dsj-2017-043