DOI: 10.1002/pra2.415

POSTERS

WILEY

Protecting participant privacy while maintaining content and context: Challenges in qualitative data De-identification and sharing

Claire A. Myers | **Shelby E. Long** | **Faye O. Polasek**

University of Michigan, Ann Arbor, Michigan

Correspondence
Claire A. Myers, University of Michigan, Ann Arbor, MI 48197.
Email: clairemy@umich.edu

Funding information
Institute of Museum and Library Services

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.

KEYWORDS

data de-identification, data sharing, qualitative research

1 | 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.

83rd Annual Meeting of the Association for Information Science & Technology October 25-29, 2020.

2 | 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 Management Expert Guide" (CESSDA Training Team, 2020), University of Michigan's guide on data management ("Data Security Guidelines," 2020), and ICPSR's suggestions on social science data management ("Guide to Social Science Data Preparation and Archiving," n.d.).

Ultimately we decided to follow more closely the De-Identification guidelines provided by QDR ("De-Identification," n.d.). More specifically we adhered to their

Qualitative Interview De-identification Protocol
Library Assessment for Research and Scholarship Lab

Purpose

To provide guidance for research teams engaged in de-identification of qualitative interview data prior to deposit and sharing.

Background

Data de-identification and sharing were included in the IRB proposal prior to beginning this research, each participant was made aware that it is the intention of the research lab to remove identifiers and share the interview data collected, and each participant signed a consent form agreeing to these terms.

The participants in this study were researchers who provided specific detailed information about their areas of study and discussed the institutions in which they work. This required specialized de-identification to address information unique to researchers, such as area of study and publications.

De-identification Process

1. Retain a copy of the original interview transcript for internal data analysis purposes.
2. Remove information identifying the interviewer and replace with corresponding number.
3. The first person doing the initial de-identification should begin by reading through the transcript for an understanding of the content and context.
4. Remove all direct and indirect identifiers and replace with more general terms while preserving meaning as much as possible. Refer to the list of identifiers and example replacement terms on pages 2 and 3 to maintain consistency between each person working on the transcripts.
5. Track all changes in a spreadsheet that is stored securely using the format below:

Interview	Page	Line	Original	Changed	Notes
P01	1	10	original text	new text	
	2	20	original text	new text	comments

6. Once the de-identification is complete and all changes are recorded in the log, a second person will review the de-identified transcript. The second person will update the transcript and log with any additional de-identification. The second person will also make any necessary updates to the formatting and style of the document to ensure consistency between all team members engaged in de-identification.

FIGURE 1 De-identification protocol, page 1

suggestions to keep a log of every alteration, develop a protocol, and document our process of creating and applying our protocol.

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-

De-identification Guide

Identifier	Description	Example	Replace with	Notes
Participant name			P01	Use participant numbers consistently across the research project and store the identifying log securely.
Interviewer name			Interviewer A	Use a unique identifier for each interviewer on the research team and store the identifying log securely.
Contact information	Phone numbers, addresses, email addresses, or other contact information of the participant or those mentioned within the interview.	I live at 101 Independence Ave.	I live at [address].	
Demographic information	The race, ethnicity, age, or gender of the participant.	I'm a white millennial woman.	I'm a [demographic information].	This can be left in if it is integral to the research question or if there is not enough information included within the complete interview to identify the participant by using this information. For example, "female astronaut" may be considered more identifiable than "female librarian."
Appearance	Description of the appearance of someone the participant describes or description of themselves.	People might recognize me because I'm tall and I have long hair.	People might recognize me because [description of appearance].	This will likely be needed relatively rarely.
Job title	The formal title of the participant.	I work as a librarian.	I work as a [occupation].	This can be left in if it is integral to the research question or if there is not enough information included within the complete interview to identify the participant by using this information.
Employer	Name or specific description of the participant's employer or employer-specific products or programs.	I work at the Library of Congress.	I work at the [institution].	It may be necessary to add context that establishes the relationship between the participant and the institution, for example [current institution name]. It may also be preferable to add clarification such as [government institution] depending on the context.

FIGURE 2 De-identification protocol, page 2

identifying and formatting the document consistently, allowing us to work at a faster pace.

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 (Figures 1–3).

Department	Description of the specific area that the participant works in within their institution.	I work in Digital Collections and digitize materials.	I work in [department] and [job description].	If the department is not specific to a particular unique organization it can be left in and any institution-specific terms can be substituted for more generic terms and phrases.
Education or work history	Mentions of the participant's education or former workplace.	I used to work at the New York Public Library.	I used to work at [prior institution].	If it is a large employer or school with many people in similar roles, it may not be necessary to remove this information, this depends on the institution.
Name of location	Description of a specific city.	I live and work in Washington, D.C.	I live and work in [city].	The context of the study will determine whether this is necessary to remove.
Name of region	Description of a specific region.	I'm from the East Coast.	I'm from [region].	This level of de-identification may not be necessary unless other identifying factors are left in that could lead to determining the participant identity.
Name of person	When the participant mentions a specific person who they know.	I work with Carla Hayden.	I work with [librarian].	It may be necessary to add context that establishes the relationship between the participant and the person mentioned, for example [librarian colleague name].
Area of study	If the participant is a researcher, writer, or other type of content creator who works in a particular area of study they may describe their unique work.	I research how librarians are impacted by student loan forgiveness programs.	I research how [specific professionals] are impacted by [federal policy] programs.	If the description of area of study is highly specific or unique, including information such as location or a particular research question they are investigating, this can be de-identified to make it more general. This information will need to be removed rarely.
Publications	Any article that the participant has published or journal that the participant has been published in.	I wrote an article titled School Librarians in School Library Journal.	I wrote an article called [title] in [academic journal].	This information could lead the reader to identify the author of a publication and should be removed if possible.
Distinct speech	Ways that participant speaks which could be considered identifiable idiosyncrasies or colloquialisms.		[...]	It is only necessary to remove if it is particularly distinct and unique enough to identify the individual speaking.

FIGURE 3 De-identification protocol, page 3

3 | 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.

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.

4 | 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.

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.

5 | 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.

ACKNOWLEDGMENTS

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

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 ed. (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. https://doi.org/10.5334/dsj-2017-043

How to cite this article: Myers CA, Long SE, Polasek FO. Protecting participant privacy while maintaining content and context: Challenges in qualitative data De-identification and sharing. *Proc Assoc Inf Sci Technol*. 2020;57:e415. https://doi.org/10.1002/pra2.415