DATA SCIENCE JOURNAL

Kirilova, D and Karcher, S 2017 Rethinking Data Sharing and Human Participant Protection in Social Science Research: Applications from the Qualitative Realm. *Data Science Journal,* 16: 43, pp. 1–7, DOI: https://doi.org/10.5334/dsj-2017-043

PRACTICE PAPER

Rethinking Data Sharing and Human Participant Protection in Social Science Research: Applications from the Qualitative Realm

Dessi Kirilova and Sebastian Karcher

Qualitative Data Repository, US
Corresponding author: Dessi Kirilova (Dessi.Kirilova@syr.edu)

While data sharing is becoming increasingly common in quantitative social inquiry, qualitative data are rarely shared. One factor inhibiting data sharing is a concern about human participant protections and privacy. Protecting the confidentiality and safety of research participants is a concern for both quantitative and qualitative researchers, but it raises specific concerns within the epistemic context of qualitative research. Thus, the applicability of emerging protection models from the quantitative realm must be carefully evaluated for application to the qualitative realm. At the same time, qualitative scholars already employ a variety of strategies for human-participant protection implicitly or informally during the research process. In this practice paper, we assess available strategies for protecting human participants and how they can be deployed. We describe a spectrum of possible data management options, such as de-identification and applying access controls, including some already employed by the Qualitative Data Repository (QDR) in tandem with its pilot depositors. Throughout the discussion, we consider the tension between modifying data or restricting access to them, and retaining their analytic value. We argue that developing explicit guidelines for sharing qualitative data generated through interaction with humans will allow scholars to address privacy concerns and increase the secondary use of their data.

Keywords: qualitative data; data sharing; sensitive data; research ethics; data curation

Introduction

While data sharing is becoming increasingly common in quantitative social inquiry, qualitative data are still rarely shared. One of the major factors inhibiting data sharing is a concern about human participant protections and privacy. Protecting the confidentiality and safety of research participants is a consideration for both quantitative and qualitative researchers, but it raises specific worries within the epistemic context of qualitative research. Thus, the applicability of emerging protection models from the quantitative realm must be carefully evaluated for elements appropriate for the qualitative realm. At the same time, qualitative scholars already employ a variety of strategies for human-participant protection implicitly or informally during the research process, so part of the challenge is lessened if data repositories help researchers draw on their familiarity and comfort with these and enhance them in the process.

In this practice paper, we draw on our experiences working at the Qualitative Data Repository (QDR) to assess available approaches for protecting human participants and how they can be deployed in the qualitative realm in particular. We describe a spectrum of possible data management options that can be used individually or in combinations, such as de-identification and applying access controls. We also review some use-case applications by the repository in tandem with its pilot depositors. Throughout the discussion, we consider the tension between modifying data or restricting access to them, and retaining their analytic value. We argue that domain data professionals, cognizant of the needs of social scientific scholarly communities, can develop explicit but flexible guidelines for sharing qualitative data generated through interaction with humans that allow scholars to address privacy concerns throughout their work process. This, in turn, will make their collected data shareable and increase their secondary use for analytical or pedagogical purposes.

Impossible to Share?

> All those records had now been burned: Even before the controversy began, Goffman felt as though their ritual incineration was the only way she could protect her friend-informers from police scrutiny after her book was published (Lewis-Kraus, 2016).

Until recently, sociologist Alice Goffman's approach to protecting her research participants was the norm in qualitative social science, even with data far less sensitive than her ethnographic study of crime and policing in low-income communities in Philadelphia (Goffman, 2014). A lack of awareness about the need for and benefits of data sharing limited practicable strategies for protecting participants, and structurally conservative institutional review boards (IRBs) all combined to dissuade researchers from even attempting to share their data. Even more fundamentally, not thinking of the qualitative materials they collect as "data" with inherent value beyond their own study, many social scientists have remained oblivious to the developing technologies, practices and scientific infrastructure that make sharing that is both legal and ethical newly possible.

The tide is turning, however: open science and research transparency are becoming established as disciplinary norms and funding agencies as well as journals are developing mandates for making articles, data, and software available to the scientific community and the public at large. Simultaneously, textual, audio, video and other types of qualitative data are becoming more immediately obtainable, and those collected in digital formats are increasingly easy to distribute. Each of these factors leads to an increased interest in managing and sharing qualitative data, but also raises concerns about how to openly share those involving human subjects both ethically and safely. The tensions between the broad vision of open access and the long-standing demand to protect the people whose information researchers use are important, but should not be declared irreconcilable. The most fruitful way forward is for institutions that fund data collection, that store data for sharing, and that publish academic work making knowledge claims on the basis of these data – in collaboration with the researchers themselves – to develop policies and procedures that are consistent with relevant legal and ethical obligations ensuring the wellbeing and privacy of research participants.

The Qualitative Data Repository (QDR, www.qdr.org), hosted by Syracuse University, went online in 2014. It was established as a domain repository with the explicit mission to provide a home for qualitative and multi-method primary data, which might otherwise remain invisible in the social science research community. QDR serves this mission most directly by offering a user-friendly platform that enables researchers from around the world and across all social science disciplines to publish their data projects in a reliable digital venue and thus make them durably discoverable (via indexing and use of digital object identifiers or DOIs), citable (by suggesting an accurate and complete bibliographical record), intelligible to others (by providing narrative documentation and structured metadata) and, ideally, linked to the original researcher's and others' publications that use them (by using CrossRef/DataCite article-data linking).

More broadly, QDR's staff – which includes the authors of this paper – has learned during these early years that its key role is to cultivate the repository's intended user community. QDR has consequently been at the forefront of efforts to promote and support the sharing of qualitative social science data, not simply by providing technical infrastructure, but by working closely with individual data depositors to curate their qualitative data for preservation and reuse and by creating useful guidance materials that address the various stages of a research project (see https://qdr.syr.edu/guidance). When provided education in the basics of data management, social scientists become well-positioned to undertake their work with the goal of sharing in mind from the planning stages. In the course of the repository operations, we have found that the biggest impact we can have is to encourage qualitative researchers to start seeing what they do as "data collection" and its artifacts as stand-alone scholarly products that are publishable and deserve intellectual recognition.

Qualitative data sharing works best when researchers are able to capitalize on their closeness to the human sources of their rich materials and on existing feelings of responsibility for and skills in protecting those sources. By giving researchers both credit for and control over their data work, we believe repositories can partner with them to advance the cause of safe and ethical data sharing. Drawing specific lessons from an initial set of pilot studies, each with its own challenges, we at QDR developed strategies to coach researchers about the options at their disposal to share even sensitive qualitative data. These strategies fit within the research data lifecycle, from planning through data publication.

Planning for Data Collection

The main insight throughout QDR's pilot projects has been the advantage of early and thorough data management planning oriented towards the later sharing of data (Karcher et al, 2016). However, many standard approaches borrowed from quantitative research are difficult to apply directly to qualitative research. For example, as a general rule of thumb, QDR recommends that scholars do not collect identifying information where it is not substantively needed for the purpose of the study. However, the nature of qualitative interviews often produces a paper (or e-mail) trail to schedule the interview where direct identifiers (names, phone numbers, addresses) abound. Complicating the situation even further, researchers often build lasting relationships spanning multiple interviews with their participants, making such a strategy inapplicable. The objection to data sharing most commonly raised by qualitative researchers themselves combines this integral role they as individuals play in the research process and the very richness of the contextual material typically gathered (Fink, 2000).

We propose to reconsider this "closeness" of the investigators, as we find that it makes them best positioned to undertake the necessary modifications to received strategies that can enable reasonable data sharing without introducing harm to the participants the researchers know so well. Instead of making the sharing and archiving of qualitative data particularly challenging, the embeddedness of researchers in their research site should be thought of as a resource, a deep foundation of knowledge of local circumstances and expectations. Thus a scholar would be able to decide in advance what might be the right secure location to store any contact information necessary for his or her ongoing interactions in the field: One example could be a notebook separate from the digital transcriptions of interviews that they keep with them at all times because of a fear that their rented apartment in the field can be accessed without their knowledge; another – a file encrypted on a memory stick, locked in a cabinet once back at their home institution, where negligence is a greater concern than unauthorized searches. Additionally, scholars who have decided on such basic data management rules in advance can use them to easily train any transcribers or other research assistants they work with in the chosen privacy protocols. Even more importantly, they can present a cogent argument during their IRB application process (i.e., before the rules are put in action) why a given option that does not involve destroying collected materials is the right choice for a given research project. Crucially, all of these downstream advantages can only be realized if the idea of sharing the data is pursued from the earliest project planning stages.

Another key aspect of qualitative data gathering concerns the informed consent procedure. As Bishop (2009, p. 261) notes, many qualitative researchers (often to accommodate what they think IRBs expect) use highly restrictive terms of consent, even where risks are minimal and research participants would not object to data sharing. Beyond requesting affirmative consent to share the collected data, researchers can and should tailor the details of their consent procedure to the locale and cultural context – and qualitative researchers can use their close interaction with participants to gain a better sense of the most appropriate form of consent. For example, we talked to one researcher studying former civil war combatants who found (somewhat to her surprise) that her interviewees were reassured by the detailed written consent forms she used. In other contexts, written consent might have the opposite effect. The guiding principle however applies to both those scenarios: the researcher needs to make intentional choices and provide clear documentation of them. Even if the decision is for verbal collective-based consent, for instance, justified on the basis of a traditional understanding of authority to grant such in the group the researcher is studying, this result and its rationale will be recorded and presented as documentation alongside the actual transcripts (full or redacted further, which should be another discrete option) of the group interviews.

Those choices themselves should be based on a thoughtful and realistic assessment of both the probability and degree of risk of harm, as weighed against the benefits of the research itself and the sharing of the data (Van Den Eynden, 2008). This sort of "risk-benefit calculation" is quite familiar to IRBs from the biomedical research realm where they originated (Beauchamp, 2011) and its logic remains broadly pertinent for social science work, both qualitative and quantitative.

Data Collection

Continuing on the topic of participants' consent, the ideal stance – which qualitative researchers should find an extension of their general ethical position with regard to the people they study – would be to involve participants in the process of data sharing. Given the extended interactions qualitative researchers typically have with their interviewees, they are able to explain the nature, purpose and benefits of data sharing and also get a sense of the types of risks their participants might be worried about (or not) and thus calibrate an initial hypothetical assessment.

What we at QDR advise researchers to do, in order to facilitate consent that grants full agency to the participants, is to offer them a range of data sharing options they can agree to. To illustrate, researchers can present separate Yes–No choices for: full audio recording; partial note taking; having one's words quoted in later research outputs, with or without attribution; and archiving, respectively, again only the transcripts or transcripts and audio recordings if both were made during the interview.[1] The customizable selections allow for a meaningful negotiation between interviewer and interviewees in a way that permits the latter to tailor their choice in a way that seems optimal to them. In all cases, participants hold a "veto" over sharing their data, but researchers should be careful to perform (together with their IRB and, later, repository staff) an individual risk assessment even where interviewees agree to data sharing.

Only careful planning before the start of the project can ensure truly informed consent (conceptualized, as we do above, as an interactive and ongoing process between researcher and participant) during the face-to-face collection stage.

Data Curation and Publication

In most cases, researchers will seek to remove personal aspects of the data for publication. There are significant challenges in de-identifying quantitative research like surveys (Kennickell, 1997), but possible attack vectors as well as possible solutions tend to be technical (e.g., adding noise to data, collapsing categories, masking or obscuring metadata records, etc.). For qualitative data, there is no alternative to a manual, context-driven procedure. While automated tools can assist by flagging possible indirect identifiers like specific dates and locations, the researcher, ideally in consultation with a data management specialist, needs to individually remove or alter numerous instances of indirect identifiers.

Some more traditional strategies for de-identification, which underlie the more recent computerized implementation, can nevertheless be more easily applied. These include broadening categories or partially reducing content. As an illustration for the first type, "I graduated from Dartmouth in 1985" becomes "I graduated [from a liberal arts college] in [1985–1990]/or [in the mid-1980s]". The resulting statement expands the population in which the interviewee falls, while also retaining critical factual information the original statement was meant to convey.

One QDR pilot project was chosen specifically to try out such a strategy for carefully de-identifying over 100 interview transcripts in a way that retained their analytical richness and inferential usefulness (Dunning and Camp, 2015). QDR curation staff worked closely with the researchers to develop a clear and comprehensive protocol of anonymization rules addressing all the detailed substantive categories that made sense in the context of this project. The resulting protocol has been archived as part of the project's documentation files (Dunning et al, 2015) and serves both as a transparent explanation of the adaptations made from the original material for a secondary user of the collection, and also as a creative model for methodological learning. A similar logic can be applied when citing such interviews in a publication. Details about interviews can be referenced either by broad categories (e.g., "city council member, Buenos Aires region, Fall 2012") or, if unfeasible, by assigning codes to the interviewees and locations (e.g., "Interviewee 1; City A").

The alternative of such content-based data redaction can be achieved by either only publishing a subset of interviews or by only publishing selected relevant extracts from them. In fact, Dunning and Camp also made the first of those additional choices, since the full data collection they and other co-authors engaged in had produced several hundred interviews. To make some data sharing practicable, while still keeping the sensitive data about political clientelism safe, the researchers selected only one geographical cluster of the several locations where interviews were conducted. Again, the choice was deliberate (random sampling across different sites could have provided similar numerical reduction, but would have eliminated the social-network aspect that was of critical importance for the research question) and clearly documented.

An interesting twist on how quantitative reduction of qualitative data can be used to protect human participants' confidentiality and safety can be seen in the so-called Active Citation compilations, commissioned by QDR (Moravcsik, 2014). In this type of project, the authors contribute annotations that supplement the arguments of a formal publication on a micro level. In some cases (Ellett, 2015; Rich, 2015), researchers who had started from the default position of no possibility of sharing, ultimately felt sufficiently comfortable to use briefer or redacted excerpts of interviews at the specific points in their texts where they were trying to

[1] An interesting recent attempt to share cases and documents, including consent scripts, from ethics reviews is The Ethics Application Repository in New Zealand. Users can use the shared materials to get ideas for their own projects (Tolich and Tumilty, 2014). QDR similarly encourages depositors to submit their approved IRB applications, in the form of a documentation file to the published project (see e.g., Flom and Post 2016).

substantiate an empirical claim. Clearly, such partial provision of data only works for this specific type of transparency technique and does not satisfy more ambitious research transparency and data access needs. But when the judgment calls of what was annotated, what type of excerpts were provided and what were excluded are openly detailed in the published project documentation, it does present a better alternative than zero primary data availability.

While the techniques discussed above all rest primarily with the researcher, another important strategy in protecting participants' privacy is not achievable without engaging a professional repository in the sharing process. This concerns the various levels of access controls, which range from simple registration requirements, to requirements to submit a research proposal for secondary data use and sign a special usage agreement, to timed embargos, or to on-site – only access. The in-depth understanding qualitative researchers have of their participants can inform the nature of access controls as well. With access options available at the file (i.e., individual interview) level, a single study can contain anything from open, identified data to unpublished data – depending on both the consent of the participant and the researcher's risk assessment. Given its flexibility, such differential access will probably become more widely used as more qualitative data are shared, as well as used, especially via institutional repositories. Restricting access might also be the only option in cases where de-identification is impossible due to the medium used, for example, for video and audio recordings. While it might be possible to blur or distort identifying elements, this is both expensive and destroys important qualities of the data.

With all these strategies, there is a trade-off between sharing and risk to privacy, between ease of access and the protection of sensitive data. QDR's goal is to teach scholars how they can reduce the trade-off to its optimal point, i.e., to share as much as possible without introducing undue additional risk. Still, just as for the research itself, where risks exist, scholars need to balance them against the numerous benefits of publicly available data. Qualitative social scientists, in particular, should do so in close collaboration with research participants, on the one end of the research process, and with domain repositories, on the other, where the staff is deeply familiar with both social science convention and qualitative methodology. QDR provides guidance for researchers throughout the research lifecycle: from planning, through handling data securely in the field, to preparing them for publication. Its published projects and training materials showcase a growing list of examples of successful data management and sharing. Researchers can apply the broader lessons learned from these materials, together with their own expertise, to arrive at context-sensitive solutions for their qualitative project.

And while some kinds of data and some ways of sharing will always be more problematic than others, often the objections to any sharing of qualitative data are based on discussing the most difficult end of the spectrum, while simultaneously envisioning the least constrained ways of sharing. While a pioneer in facilitating the sharing of qualitative data, QDR does not advocate such an imaginary "handing over the data" (Sieber, 1988) without sufficient preparation. Various tools for advance planning and constrained sharing, in the cases where that is appropriate, should allow even the more difficult cases to be handled properly with great benefits to the scholarly enterprise and to individual research projects. If research is conducted with eventual sharing in mind, and if scholars are familiar with the available strategies, then the kinds of dilemmas that for a long time have forced researchers into promises for data destruction become the exceptions and no longer the rule.

A Cautionary Ending (for would-be data incinerators)

> Earlier that day, I'd taken a train from New York to Philadelphia because I wanted to track down at least one of Goffman's subjects, and I was pretty sure I had figured out where some of Chuck's surviving family lived (Singal, 2015).

While Alice Goffman's motivation to protect her Philadelphian interlocutors was certainly admirable, the way she went about it seems to have caused a lot of negative scrutiny of her work and, most unfortunately, not a lot of protection in the end. As a new institution trying to learn from the theoretical advances and good practices of archivists and information scientists, QDR's current prescriptions rest on trying to maximize those of the so-called "Five Safes" (Corti and Welpton, 2015) that are and will remain most relevant for qualitative work in the social sciences. While "safe data" and "safe outputs" can rarely be produced from sensitive qualitative materials, educating researchers how to be "safe people" and how to plan for "safe projects" – when accessing such data and using them for secondary analysis – and providing long-term "safe settings" for the data, including via de-identification and appropriate access controls, will remain QDR's focus in its future work with researchers.

Going forward, QDR will continue to build upon the lessons referenced above from its early deposits by pursuing interactions with all relevant actors along various avenues in the social science enterprise. We are currently distilling key insights from our first three years of operations into an expanding suite of guidance documents. QDR continues to improve its technical platform, in order to offer researchers fine-grained control over and easy application of access controls. To further facilitate sharing of sensitive data, we are conducting a multi-prong outreach effort to the U.S. IRB community. By building relationships between repositories and IRBs, we hope to improve the review of data-sharing provisions in IRB applications and assure that data sharing is not unecessarily impeded by IRB protocols. Most importantly, the repository's primary enduring commitment remains to future depositors and the participants in their projects, to work together to present creative solutions for qualitative data management and sharing in a way that is both ethical and productive.

Competing Interests

Both authors work at the Qualitative Data Repository, which is discussed in detail in the article. They do not have any other competing interests.

References

Beauchamp, T L 2011 Informed Consent: Its History, Meaning, and Present Challenges. *Cambridge Quarterly of Healthcare Ethics*, 20(04): 515–523. DOI: https://doi.org/10.1017/S0963180111000259

Bishop, L 2009 Ethical Sharing and Reuse of Qualitative Data. *Australian Journal of Social Issues*, 44(3): 255–272. DOI: https://doi.org/10.1002/j.1839-4655.2009.tb00145.x

Corti, L and **Welpton, R** 2015 Access to sensitive data for research: "The 5 Safes." *Data Impact Blog*. Available at: http://blog.ukdataservice.ac.uk/access-to-sensitive-data-for-research-the-5-safes/ [Last accessed 7 June 2017].

Dunning, T and **Camp, E** 2015 Brokers, Voters, and Clientelism: The Puzzle of Distributive Politics. Data Collection, QDR:10055. Syracuse, NY: Qualitative Data Repository [distributor]. DOI: https://doi.org/10.5064/F6Z60KZB

Dunning, T, Camp, E and **Cecchi, L** 2015 Brokers, Voters, and Clientelism – Anonymization Protocol Documentation. QDR:30567. Syracuse, NY: Qualitative Data Repository [distributor]. DOI: https://doi.org/10.5064/F6Z60KZB

Ellett, R 2015 Data for: Democratic and Judicial Stagnation. Active Citation Compilation, QDR:10064. Syracuse, NY: Qualitative Data Repository [distributor]. DOI: https://doi.org/10.5064/F6PN93H4

Fink, A S 2000 The Role of the Researcher in the Qualitative Research Process: A Potential Barrier to Archiving Qualitative Data. *Forum: Qualitative Sozialforschung/Qualitative Social Research*, 1(3): Art. 4. http://nbn-resolving.de/urn:nbn:de:0114-fqs000344.

Flom, H and **Post, A E** 2016 Data for: Blame Avoidance and Policy Stability in Developing Democracies: The Politics of Public Security in Buenos Aires. Data Collection, QDR:10068. Syracuse, NY: Qualitative Data Repository [distributor]. DOI: https://doi.org/10.5064/F6RF5RZV

Goffman, A 2014 *On the run: fugitive life in an American city*. Chicago and London: The University of Chicago Press. DOI: https://doi.org/10.7208/chicago/9780226136851.001.0001

Karcher, S, Kirilova, D and **Weber, N** 2016 Beyond the matrix: Repository services for qualitative data. *IFLA Journal*, 42(4). DOI: https://doi.org/10.1177/0340035216672870

Kennickell, A B 1997 Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In: *Record Linkage Techniques: Proceedings of an International Workshop and Exposition*. Presented at the Record Linkage Techniques. Arlington, VA: Federal Committee on Statistical Methodology, Office of Management and Budget, 248–267.

Lewis-Kraus, G 2016 The Changeling: The Trials of Alice Goffman. *The New York Times Magazine*, 17 January. Available at: http://www.nytimes.com/2016/01/17/magazine/the-trials-of-alice-goffman.html.

Moravcsik, A 2014 Transparency: The Revolution in Qualitative Research. *PS: Political Science & Politics*, 47: 48–53. DOI: https://doi.org/10.1017/S1049096513001789

Rich, J 2015 Grassroots Bureaucracy: Intergovernmental Relations and Popular Mobilization in Brazil's AIDS Policy Sector. Active Citation Compilation, QDR:10066. Syracuse, NY: Qualitative Data Repository [distributor]. DOI: https://doi.org/10.5064/F6SF2T3N

Sieber, J E 1988 Data Sharing: Defining Problems and Seeking Solutions. *Law and Human Behavior*, 12(2): 199–206. DOI: https://doi.org/10.1007/BF01073128

Singal, J 2015 The Internet Accused Alice Goffman of Faking Details in Her Study of a Black Neighborhood. I Went to Philadelphia to Check. *NY Mag – Science of Us*. Available at: http://nymag.com/scienceofus/2015/06/i-fact-checked-alice-goffman-with-her-subjects.html [Last accessed 9 June 2017].

Tolich, M and **Tumilty, E** 2014 Making ethics review a learning institution: The Ethics Application Repository proof of concept. *Qualitative Research*, 14(2): 201–212. DOI: https://doi.org/10.1177/1468794112468476

Van Den Eynden, V 2008 Sharing Research Data and Confidentiality: Restrictions Caused by Deficient Consent Forms. *Research Ethics*, 4(1): 37–38. DOI: https://doi.org/10.1177/174701610800400111

How to cite this article: Kirilova, D and Karcher, S 2017 Rethinking Data Sharing and Human Participant Protection in Social Science Research: Applications from the Qualitative Realm. *Data Science Journal,* 16: 43, pp. 1–7, DOI: https://doi.org/10.5334/dsj-2017-043

Submitted: 01 November 2016 **Accepted:** 25 July 2017 **Published:** 07 September 2017

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS