Open access to qualitative research data : A guide for PhD students in the social sciences at the University of Helsinki

Khan, Jawaria

University of Helsinki, Faculty of Social Sciences

2021-12

Khan , J , Rantanen , V E T & Kolmonen , P T 2021 ' Open access to qualitative research data : A guide for PhD students in the social sciences at the University of Helsinki ' Population ageing and its social consequences working paper , no. 9 , University of Helsinki, Faculty of Social Sciences , Helsinki .

http://hdl.handle.net/10138/337870

unspecified
publishedVersion

UNIVERSITY OF HELSINKI
FACULTY OF SOCIAL SCIENCES

Jawaria Khan[1], Visa Rantanen[2], Petteri T Kolmonen[3]

Open access to qualitative research data: A guide for PhD students in the social sciences at the University of Helsinki

Population ageing and its social consequences working paper No. 9
Published December 2021
Place: Helsinki, Finland

Abstract

This study emphasises the importance of preserving qualitative research data for secondary use. Universities and funding agencies ask researchers to consider data sharing as part of their research design and funding proposals. There has been significant growth in making data openly available for reuse. Unlike quantitative data, qualitative data is unstructured, and it is difficult to give access to sensitive information contained in the data. Thus, this paper is a source of help for qualitative researchers in multiple ways. First, it provides an analysis of open data science and its process through information provided in University of Helsinki's webinars and training courses. Second, it answers the concerns of doctoral students dealing with qualitative data in the light of existing literature. Third, it concludes by offering some useful tips for making the data openly accessible to researchers dealing with qualitative data.

Keywords: open access data; qualitative data; open data science; data publishing

1. Introduction

Science is made up of data, its collection, analysis, use, reuse, share and reshare (Molloy, 2011). However, this information dissemination gets stuck when the scientific data are not made available to be used again. The data remains 'confidential' and there is a personal reluctance to openly publish the data especially qualitative data in Social Sciences. One of the reasons for this reluctance is the fear of releasing the data 'into the wild' (Molloy, 2011). The fear is that the data could be used improperly, inaccurately and the owner of the data might not get credit or any incentive to make the data openly available. The solution to these fears about accessing the scientific data by researchers have been addressed through Open Data science.

Open access to data refers to the process through which the scientific data can be published and openly re-used free of charge (Murray-Rust, 2008). It means that collected and retained data can be used again after ethical considerations. The advantages of granting access to data for researchers could be to 'replicate, verify and expand' their scientific research (Andreoli-Versbach, & Mueller-Langer, 2014). For the scientific research, the benefits of releasing the data comprises of reduced corruption related to faulty data and detection of inaccuracy in the data.

Unlike quantitative data, qualitative data are much difficult to contextualise adequately, and it becomes challenging to use the decontextualised data (Chauvette, Schick-Makaroff & Molzahn, 2019). However, sometimes qualitative data can be used scientifically while decontextualised, for example a linguist can study historical changes in speed patterns using archives of past qualitative interviews if they are adequately ad verbatim transcribed, regardless of what the actual interviews were about.

The concerns of qualitative researchers have not been discussed in the context of opening their research data. Hence, the purpose of this paper is threefold. First, it provides an analysis of the open data science and its process through information provided in University of Helsinki's webinars and training courses. Second, it answers the concerns of doctoral students dealing with qualitative data in the light of existing literature. Third, it concludes with offering some useful tips for making the data openly accessible to researchers dealing with qualitative data.

2. OPEN DATA SCIENCE

Open access to research data does not mean or demand that everything be opened up but guides the researchers on basis under which the data can be made openly available. The European

Commission has aptly put this slogan for open data: ***"As open as possible, as closed as necessary".*** AILA, the Finnish Social Science Data Archive reports the conditions for open access of a dataset (see *Table 1)*:

Table 1. Conditions for open access of a dataset

A	openly available for all users without registration (CC BY 4.0),
B	available for research, teaching and study,
C	available for research only (including master's, doctoral and polytechnic/university of applied sciences master's theses),
D	available only by permission from the data depositor/creator.

Note. From Finnish Social Science Data Archive. (n.d). (https://services.fsd.tuni.fi/catalogue/index?lang=en&study_language=en). CC BY.

Open access to research data means that data produced in one project can be reused again as 'raw data' in another. The data can be duplicated, enhanced and developed from one project to another at any stage: from raw material to data, from data to analysis, from analysis to interpretation and from interpretation to results. To open the research data, a researcher requires a huge amount of planning early in the collection phase. To avoid complex issues, the researchers should plan at the beginning about the quality, consistency, validity, responsibility and anonymity of data through a data management plan. In addition, funding agencies have also some set terms about opening access to the data.

The concept of open data was introduced by the Organisation for Economic Cooperation and Development (OECD, 2007) in the United Kingdom during the mid-1990s on the principles of openness, flexibility, transparency, accountability, legal and ethical compliance and free exchange of information. Several OECD member countries including Finland have also followed the rules of open data directed by OECD. It has been more than two decades since the world's social science qualitative data archive was first established by the United Kingdom's Qualidata initiative (Corti et al., 2014). It was followed by other data archives in the US, Australia and Europe. The Finnish Social Science Data Archive was established in 2003 with the aim of focusing on qualitative data archives too. This movement towards open data is significant because with time there is an increase in demand for open data access by the funding organisations for researchers.

2.1 How can it be done?

Core Trust Seal is a certification organisation that catalogues trustworthy data repositories that allow for data sharing. Certification may be granted to a data repository that fulfils all requirements, all of which are equally important and evaluated thoroughly as stand-alone items. The repositories that qualify for Core Trust Seal certification are trustworthy archives of data to be made accessible to the scholarly community.

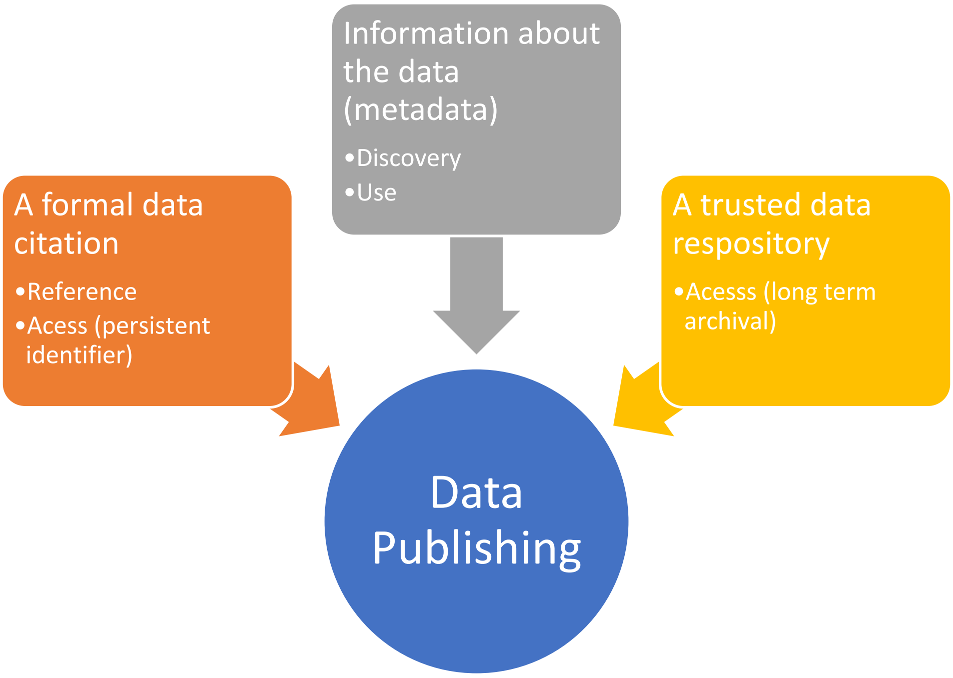
In Finland there are two Core Trust Seal certified data archives:

- The Language Bank of Finland, which focuses on storing spoken and written language material.
    - Guidelines for data submissions (in Finnish): https://www.kielipankki.fi/tuki/ohjeita-sisallontuottajille/
- The Finnish Social Science Data Archive, which stores both quantitative and qualitative data.
    - Guidelines for data submissions: https://www.fsd.tuni.fi/en/services/depositing-data/guidelines-for-depositing-data/

The data submitted to a data archive must comply with the guidelines that ensure the trustworthiness and the ethical use of the data. These guidelines typically include specifying the full contents of the data, the requirements for anonymisation of the data, a full description of the data collection process, an agreement on the processing of the data, ensuring that the technical format of the data is correct, and other requirements which may be specific to each archive in question. This description of one's own data like time, location, data method, file names, file formats, versions, variables etc. is called metadata. Hence, it is advisable for all the collaborators to follow a consistent file naming system. Also, it is recommended having README files with the original data which provide information about data files to understand them.

Once the data have been submitted to the archive, it will be assessed for suitability and if approved, it will become available for users to download for their use. The case for using the data may be set in accordance with how one wishes the data to be made available. It can be possible to make one's data available to all users, or restrictions can be applied on who can access the data, for what purpose, or if the permission of the data holder is needed for accessing the data. One must note here that data archiving is different from data publishing. Data archiving is the long-term storage of data and methods not necessarily opened; however, data publishing is the about the opening the data *(see Figure 1)*. But both require metadata.

Figure 1. The process of data publishing


Information about the data (metadata)
•Discovery
•Use
A formal data citation
•Reference
•Acess (persistent identifier)
A trusted data respository
•Acesss (long term archival)
Data Publishing

Note. From RDM basics, University of Helsinki Data Support, Spring 2020, (https://moodle.helsinki.fi/mod/resource/view.php?id=1873246). CC BY.

2.2 What are the challenges of open data?

The following are the recorded challenges of opening the data public access.

2.2.1 How to ensure the consistency and quality of data?

The researchers can tackle the issues of consistency and quality by thoroughly organising the data collection phase of their research through a data collection plan. It will allow them to have transparent and accountable processes at every stage of data collection and recording. The methods of ensuring that one's data maintains the original and accurate information about sources include proper and accurate recording and representation of data, peer review of data and calibration of data. The quality assurance steps can be documented in the metadata and can be consulted through Openscience.fi. Usually, the methods of ensuring quality are termed as FAIR: Findable, Accessible, Interoperable and Re-usable (Fairdata.fi) see Table 2 for reference.

Table 2: FAIR principles and definitions

Principle	Definition
Findable	• Data are described with rich metadata • Data or metadata have unique and persistent identifier • Metadata includes the identifier of the data it describes • Data or metadata are registered/indexed in searchable resource
Accessible	• Data or metadata are retrievable by identifier using a standardised protocol • Protocol is open, free and universally implementable • Protocol allows for authentication and/or authorisation when needed • Metadata remain accessible even if data are no longer available
Interoperable	• Data and metadata use a formal, accessible, shared and applicable language to represent content and knowledge • Data and metadata use vocabularies that follow FAIR principles • Data and metadata include and describe references to other data/metadata sources
Re-usable	•Data and metadata have rich description and plurality of accurate and relevant attributes • Data and metadata have clear and accessible usage license • Data and metadata include detailed provenance information • Data and metadata meet domain-relevant community standards

Note. From "*Qualitative data sharing and re-use for socio-environmental systems research: A synthesis of opportunities, challenges, resources and approaches*," by Jones et al., 2018, SESYNC White Paper, p. 9 (10.13016/M2WH2DG59)

2.2.2 Who maintains the rights of ownership?

The rights of ownership of research data depends on the research funding. If the research is funded by the university as open research, then the project ownership rights remain with the researcher, but in case of external funding, researchers must make an agreement on the transfer of rights to their university. The researchers are not permitted to transfer the unpublished data outside the university without contacting University of Helsinki Legal researchlawyers@helsinki.fi for agreements on confidentiality and reuse while being affiliated with the University of Helsinki.

2.3.4 Who has the rights of usage of one's data?

The usage rights depend on the individual or group which collects the data. It is advisable in group projects for the principal investigator or any other researcher to be assigned the right of issuing reuse at the beginning of data collection phase. Also, the publishers are advised to give proper credit to the original researchers as well as those who do secondary analysis (Jones et al., 2018)

3 CONCERNS OF A PhD STUDENT WORKING WITH QUALITATIVE DATA

Unlike quantitative data sets, qualitative data are of several types. The data can consist of interview recordings (audio/video), text file (transcripts), focus group discussions, meeting minutes, a personal diary, oral history interviews, field notes, scans of newspaper articles, images of official documents, policy documents, pictures, videos, emails etc. This kind of data is considered to be unstructured, and it is challenging to organise and publish it. Generating research data usually involves creating secondary data in addition to the primary research data. Additional data can include lists of research participants and their personal details. Research can also create audio- and video recordings. These can be used for research purposes, but often they are treated merely as instruments for creating anonymised transcripts. Hence, there are some concerns or questions from the perspective of a doctoral student dealing with qualitative data which are answered in the section below.

3.1. What about the role of researcher-participant relationship?

In qualitative research settings, participants have a major role to play in the research process. Some researchers argue that the quality of the data produced sometimes depends on the relationship in terms of interaction between the researcher and the participants (Broom et al., 2009; Carusi and Jirotka, 2009). Bishop (2009) explains that this issue of relationship is all about how much importance is given to research subjects in a study starting from informed consent to the implications of the results.

3.3 How is qualitative data re-used?

Corti (2000) describes six ways in which qualitative datasets can be reused. First, having new research questions for old data and approaching the data in a way that was not done in original project. Second, the data can be used as the research design of new study. Third, the data can be used for teaching and training in research methods. Fourth, for comparative studies, to compare old data with new data across time, region, groups, etc., Fifth, for the verification of the results. Finally, the published data will be historical resource.

3.3. What about informed consent?

Informed consent ensures the legal and ethical compliance of research for the participants. General Data Protection Guidelines (GDPR, 2021) make researchers ethically responsible for the protection of the rights of their participants in terms of confidentiality and anonymity.

Therefore, there are legal and ethical concerns about sharing the data in qualitative studies. It is the right of the participants to know about how their data will be used and reused. However, research suggests that it is impossible to guarantee participants about the reuse of data (Chauvette, Schick-Makaroff & Molzahn, 2019; Bishop, 2009). Consent is also more complicated with ethnographic data collection than with structured interviews.

Heaton (2008) suggests getting 'blanket consent' which enables the researchers to use the data indefinitely and use and reuse them for any purpose. However, it might be in conflict with the GDPR provisions of informed consent requirements which requires clarity about one's intentions on how the data will be used. Hence, it has been made explicit in the GDPR guidelines (Wolford, 2019):

> *There's no question the GDPR makes it more difficult to profit from other people's personal data. But that's the point of the law: it's other people's data; if you want to use it, you need to have a good reason, or just ask.*

3.4 What about the data collected from participatory research?

Participatory research raises certain concerns about making the data available for reuse because the data comprise 'lived experience' and field notes in the form of personal diary and headnotes. It means that participatory research does not capture all the data in transcripts. It is also about both the verbal and non-verbal experiences like body language, the reactions and situatedness of the participants and the experience, impressions and epiphanies of the researcher. Thus, it should be noted that not everything the researcher sees, feels, hears, observes can be written down. The participants are active contributors to the research process. Therefore, the use of field notes for reuse as a data source is challenging because they can lead to misunderstanding and be misinterpreted (Chauvette, Schick-Makaroff & Molzahn, 2019). In essence, ethnographic notes, which are research data, can contain information that is hard to understand without being the researcher and some information pertaining to the researcher's relationships to the observed subjects that can be too personal to be shared, at least in raw form. However, if they are edited for the wider public, they stop being data and become something else. Thus, this issue is complicated.

3.5 What about the reflexivity?

The essence of qualitative ethnographic research lies in the researcher 'being there', something which cannot be shared. Being there also encourages reflexivity and ongoing self-critique in the field work through which researcher's feelings are exposed. Reflexivity is important for assessing the researchers subjective influence on the data that is collected. Thus, a researcher's reflexivity helps in collecting, storing and reusing data in an unbiased, original and rigorous way.

3.6 What about personal incentives for voluntary data sharing?

The incentive structure of publish or perish promotes the sharing and reproduction of data. It saves the time and resources of a researcher. For this reason, research suggests that non tenure track workers have a higher incentive not to publish their data openly than the tenure track ones because it is of more value to them (Haeussler et al., 2014). However, the institutions and funders have made open access to research data a requirement for funding applications and research publications (Bishop, 2009). Jones et al. (2018) explain the material incentive to publish open data in an apt way:

> *The ability to further learn from and interpret secondary qualitative data is especially important for early-career researchers and those not situated in academic institutions, for whom securing governmental funding is more challenging, as well as for practitioners with methodological training who sit outside of traditional research institutions but have the interest and ability to use qualitative analyses to inform their work* (p. 4).

3.7 What about non-anonymous data?

Typical open access data includes anonymous numerical data or transcribed texts. But it is often useful for the data collecting researcher to hold on to the recordings and participants' personal data. There may be a need to consult original recordings in cases of transcription error. Original research participants might have to be contacted to confirm something that the primary data leaves unclear. This creates alternative subsets of research data, some of which are hard to anonymise meaningfully and some of which are not meant to be fully anonymised. These subsets have different rules and life spans for archiving and sharing.

Records that identify research participants are generally kept for at least the duration of the research project, but sometimes they are kept longer. Since the data cannot be shared, it might be useful for retention in a non-public archive, from which the data can be accessed if needed. Preserving these data entail certain requirements. The European General Data Protection Regulation (GDPR.eu, 2021) gives people the right to be informed on what personal details organisations have on them and to have their data destroyed upon request. An organisation must name a person who acts as the controller responsible for the personal data kept in the archives. If this requirement cannot be met, it might be a reason for destroying the dataset containing research participants' personal data. Recordings from which the research participants can be identified are usually destroyed instead of being shared at the end of the research project, unless retaining and sharing the data has been agreed on with the research participants at the start of the data collection (Kuula, 2006). The process of data collection, personal data control, anonymisation, and data sharing should be planned with the research stakeholders at the start of the research project.

Scientists seek to maintain open access data as long as possible. Collecting social science data consumes time and resources. Therefore, it is important to maximise the use of the primary research data. Anonymisation is usually seen as a sufficient practice for harmonising this goal with the privacy and security of the research participants. Fully anonymised data are also not regarded as being personal data. This ethical approach is usually followed in social science data archives.

There can be additional rules for divergent cases, in which personal identifying information is shared alongside the rest of the data. For example, in the oral history tradition, research data may be seen as necessary for understanding the record of historical events. The standpoint and position of the individual interviewee in relation to the events is important for assessing the credibility of the data, therefore the respondent's personally identifiable information may be regarded as important as the rest of the data. This requires consent from the research participants for their identities to be shared alongside the rest of the data (Parry & Mauthner, 2004). In qualitative sociological research, data are collected to generate research findings and new theories. Theories imply some level of universality or transcendence. The personhood of the research participant is not necessarily important for evaluating the data as such, which is why the data are still useful even when anonymity removes it from its creation context.

3.7 Are there any access levels in open access?

Since the principle of open data focuses on 'as open as possible and as closed as necessary', the data can be categorised according to several levels of restrictions. For instance, Jones et al., (2018) have provided the following information (see Table 3) about level of access:

Table 3: Levels of data access with few examples

Level of Access	Definition	Possible Examples of Qualitative Data
A. Open	Data are freely available for use in accordance with general use agreement of repository and standard citation practices	Public policy documents, images from a political event with blurred faces, thematic analysis of interviews
B. Restricted	Data are available for use when user meets standard criteria set by the data repository to ensure ethical use of data	Written summary of sensitive data with reference
C. Controlled	Data are available for use when the user is approved by the original researcher (access could depend on research questions and intended analysis, access method and amount of data shared is decided by the original researcher)	Excerpts of ethnographic field notes, interview transcripts with names and sensitive information, raw interview data
D. Closed	Data deposit and citation exist for archival purposes, but no data are currently available (could be embargoed until publication of results, change in sensitive situation, death of a participant, or certain duration of time from collection)	Photographs of sensitive sites or individuals

Note. Adapted from "*Qualitative data sharing and re-use for socio-environmental systems research: A synthesis of opportunities, challenges, resources and approaches*," by Jones et al., 2018, SESYNC White Paper (10.13016/M2WH2DG59)

4. CONCLUSION

Two decades ago, researchers would not have thought about sharing their data and allowing it to be reused. However, due to institutional support and the demands of funding agencies, now there is much more emphasis on the ethical sharing of the data. In a nutshell, there are several advantages of making data publicly available through verified archives. *Table 4* summarises the concerns of doctoral students dealing with qualitative data sets and provides some tips.

Table 4. Tips for PhD students dealing with qualitative datasets

Concerns of PhD student dealing with qualitative data	**Useful tips**
Why should I have open access?	• institutional and funding agencies demand • academic credit and more visibility • more citations of original publications • verification of research findings • improvement in research methods • important source of teaching and learning
What about sensitive data?	Ask for consent from participants at the beginning for sharing; anonymise the data to protect identity; and use controlled access (Corti et al., 2014). Remember: "as open as possible, as closed as necessary"
What if, I have not asked for consent to reshare data at the beginning?	You can always seek retrospective permission from participants. However, they are still free to consent or not (Corti et al., 2014).
What about audio-visual data?	The video can be blurred, and voices can be distorted.
What if I promised to destroy the data once the project ends?	First, avoid making such promises. Negotiate with your research ethics council, institutional review board and funder about this agreement (Corti et al., 2014).
Why should I opt for/use secondary data from open data resources?	It saves resources and it is economical. Also, it will avoid repetitive analytical frames.
When to do it? (before your publications or after)	It is your choice; you can publish raw data, or you can publish the data after the project results have been made public.
What could the checklist be if I plan to publish my data?	• Make deals about ownership of data in good time through a data management plan (DMP) • Think about where to publish the data during the planning stage • Consider the requirement of the data archive to documentation, metadata and file formats • Remember the data protection of participants • Choose a suitable license to your data

Acknowledgement

This working paper has been supported by a grant from the Helsinki Institute for Social Sciences and Humanities (HSSH) for the project 'Sharing data in the social sciences'.

References

Andreoli-Versbach, P., & Mueller-Langer, F. (2014). Open access to data: An ideal professed but not practiced. *Research Policy, 43*(9), 1621-1633. https://doi.org/10.1016/j.respol.2014.04.008

Bishop., L. (2009). Ethical sharing and reuse of qualitative data. *The Australian Journal of Social Issues, 44*(3), 255-272. https://doi.org/10.1002/j.1839-4655.2009.tb00145.x

Chauvette, A., Schick-Makaroff, K., & Molzahn, A. E. (2019). Open data in qualitative research. *International Journal of Qualitative Methods, 18*, 1609406918823863

Corti, L. (2000). Progress and problems of preserving and providing access to qualitative data for social research—the international picture of an emerging culture. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 1*(3). https://doi.org/10.17169/fqs-1.3.1019

Corti, L., Witzel, A., Bishop, L., & Woollard, M. (2014). *Managing and sharing research data. A guide to good practice.* Thousand Oaks, Sage Publications.

Finnish Social Science Data Archive FSD.FI. (n.d.). Retrieved November 20, 2021, from https://services.fsd.tuni.fi/catalogue/index?lang=en&study_language=en). CC BY.

GDPR.EU (n.d.). *Complete guide to GDPR compliance.* Retrieved October 26, 2021, from https://gdpr.eu/

Haeussler, C., Jiang, L., Thursby, J., & Thursby, M. (2014). Specific and general information sharing among competing academic researchers. *Research Policy, 43*(3), 465-475.

Heaton, J. (2008). Secondary analysis of qualitative data: An overview. *Historical Social Research, 33*(3), 33-45.

Jones, K., Alexander, S., M., Bennett, N., Bishop, L., Budden, A., Cox, M., Crosas, M., Game, E., Geary, J., Hahn, C., Hardy, D., Johnson, J., Karcher, S., LaFevor, M., Motzer, N., Pinto da Silva, P., Pittman, J., Randell, H., Silva, J., Smith, J., Smorul, M., Strasser, C., Strawhacker, C., Stuhl, A., Weber, N., & Winslow, D. (2018). *Qualitative data sharing and re-use for socio-environmental systems research: A synthesis of opportunities, challenges, resources and approaches.* SESYNC White Paper. DOI:10.13016/M2WH2DG59.

Kuula, A. (2006) *Tutkimusetiikka: Aineistojen hankinta, käyttö ja säilytys.* Tampere, Vastapaino.

Molloy, J.C. (2011). The open knowledge foundation: open data means better science. *PLoS Biology, 9*(12): e1001195. https://doi.org/10.1371/journal.pbio.1001195

Murray-Rust, P. (2008). Open data in science. *Serials Review, 34*(1), 52-64. https://doi.org/10.1016/j.serrev.2008.01.001

Organization for Economic Cooperation and Development. (2007). *OECD principles and guidelines for access to research data from public funding.* https://www.oecd.org/sti/inno/38500813.pdf

Parry, O., & Mauthner, N., S. (2004). Whose data are they anyway? Practical, legal and ethical issues in archiving qualitative research data. *Sociology, 38*(1), 139-152.

RDM basics. (Spring 2020). University of Helsinki Data Support Retrieved November 28, 2021 from https://moodle.helsinki.fi/mod/resource/view.php?id=1873246

Wolford, B. (2019). Data sharing and GDPR compliance: Bounty UK shows what not to do. Retrieved October 30, 2021, from https://gdpr.eu/data-sharing-bounty-fine/