Qualitative Data Primer

Authors: Diana Castillo, Heather Coates, Mikala Narlock

Mentor: Shanda Hunt

About: Curating datasets underlying qualitative research

Related DCN Primers: ATLAS.ti Primer, Nvivo Primer, Human Subjects Data Essentials Primer

Topic	Description
What is qualitative data?	Qualitative data can be wide-ranging, from the more traditional surveys and interviews, to photos and other images, to social media posts (Flick, 2014). The goal of analyzing qualitative data is to explore the experiences, interactions, and related materials to trace the relationships and help describe the phenomena being studied (Flick, 2018).
Structure	Projects are often exported as packages of files to be loaded into other qualitative data software.
Primary fields or areas of use	Qualitative data is used by many different disciplines for a wide-variety of reasons, ranging from literature review to analyzing themes in interviews, datasets, images, and audiovisual materials. Common disciplines tend to be the social sciences, but many different fields could use these tools for qualitative or mixed methods research. -Humanities -Social Sciences: Political Science, Sociology -Health Sciences: Public Health, Nursing, Health Services -Qualitative or mixed-method researchers
Source and affiliation	There are open-source and proprietary software options. Proprietary: Nvivo (software) ATLAS.ti (software) Dedoose (browser based)

	Open Source: qcoder (an R library for qualitative analysis of text) RQDA (an R package for qualitative analysis of plain text) Taguette (open source qualitative analysis program that works on Windows, Mac, and Linux computers, as well as in-browser)
Software exchange standards	The REFI-QDA Standard : <https://www.qdasoftware.org/ QuDEX - The Qualitative Data Exchange Schema (QuDEx). Used by UK Data Archive DDI Standard: <pdf>[1]
Key questions for curation review	● What is the minimal level of documentation (e.g., codebook, node structure, etc.) required to accept this deposit in your repository? ● Which software was used? Which version of the software? Has this been recorded in the readme? ● Has the file been exported in a proprietary format or open source (e.g., in ATLAS.ti, one can export as an atlproj or an open source qualitative data format)? ● Which types of source formats were used for analysis (e.g., text, image, A/V, etc.)? ● Are there any potentially sensitive or protected data, including personally identifiable information? ● If the data have been anonymized, is there a record of what has been redacted or changed? ● Does the project include associated information, including codebook(s)? ● What contextual information about the coding and analytical process is crucial for the evaluation and interpretation of the published findings?
Tools for curation review	Original software used to create the project, spreadsheet or text editor.
Date Created	October 2020
Created by	Diana Castillo, Heather Coates, Mikala Narlock
Date updated and summary of changes made	March 11, 2021 published
Suggested Citation	Castillo, Diana; Coates, Heather; Narlock, Mikala. (2021). Qualitative Data Curation Primer. Data Curation Network. Retrieved from the University of Minnesota Digital Conservancy, https://hdl.handle.net/11299/219053.

[1] ATLAS.ti primer

Table of Contents

Scope

This qualitative data primer is designed to give data curators a grounding in what goes into evaluating qualitative datasets. This primer is software agnostic, focusing specifically on the specific needs of qualitative data curation.

What is Qualitative Data?

The definition of what counts as qualitative data has changed over time, but can be understood as empirical data that cannot be expressed numerically without losing important contextual information. This can include surveys, interviews, photographs, audiovisual materials, and social media posts. Qualitative research is conducted in numerous disciplines, and is often generated by mixed-methods design. Qualitative data lends itself to more holistic analysis than does quantitative data.

Files needed for curation of a qualitative dataset

Necessary

- Original source data, when possible, and as long as it is de-identified
- Processed data (e.g., transcripts)
- Codebook
- Readme
- Instrument for data collection (e.g., survey, interview, etc), if applicable
- Documentation generated in the Qualitative Data Analysis Software (QDAS) application (e.g., Nvivo, atlas.TI, etc.): memos, notes, networks, classifications

Ideally:

- Informed consent statement(s) or assent (see the Human Subjects Data Essentials Primer)
- IRB protocol (see the Human Subjects Data Essentials Primer)
- Study protocol or procedures manual
- Data Management Plan --To indicate how long the data needs to be preserved/accessible, and any other stipulations on sharing/management.

Ensure the files are present, and can be opened and understood. If possible, save the data in both the project-specific file type (e.g., .atlproj for ATLAS.ti) and an export of the raw data in a csv, text, or other open source file(s). Some software allow users to export the entire project in an open source format, such as a csv.

Selecting an appropriate data repository

When selecting a repository, there are many factors to consider before depositing your data and content. Every repository has a different mission and will comply with different requirements. If your data contains personally identifiable information (PII) or is subject to legal or regulatory requirements that will require controlled access to data and/or research outputs, it is critical that a repository is evaluated before the data is created.

This is not an exhaustive list of considerations, but every potential repository should be evaluated on the most pressing needs of the dataset, such as:

- Data sensitivity -- How sensitive is the data? What protections should be in place? Can the repository provide the appropriate level of protection to meet relevant legal or regulatory requirements?
- Persistence -- Will the repository be able to provide a persistent and unique identifier for your data?
- Preservation -- Some repositories are preservation focused and may require strict compliance with their regulations. Are there any format restrictions on what the repository will accept? Are there strict metadata standards you must comply with? If preservation is crucial, do they have a contingency plan in the event they must suspend service? Have they received certification from any national or international agencies?
- Costs -- Many repositories have a cost-recovery model in place for deposits, particularly for large datasets. Does the repository charge for upload or maintenance of the data? If so, is there funding available to cover these expenses?
- Access -- Can your data be made openly accessible? Do you need to restrict access to part or all of the data? Do you need to add collaborators as “Editors” to the data? Do you need to have usage statistics tracked? If so, can the repository provide that information?

In addition to the above requirements, consider how the data and any publication records, such as presentations and articles, will be linked together regardless of where the content is stored. If the publisher will store both the data and the publication, consider storing a backup copy of the data in another repository that is publicly available, such as an institutional repository.

Steps & Tools for Reuse of Qualitative Data

Moravcsik (2014), a political scientist, describes three dimensions of research transparency:

- Production transparency, which grants readers access to information about the methods by which particular bodies of cited evidence, arguments, and methods were selected from among the full body of possible choices.
    - Strategies include a Data Management Plan, documentation throughout the study, and organizing with the goal of sharing and reuse (even if by you alone).
- Analytic transparency, which assures readers access to information about data analysis. This entails the precise interpretive process by which an author infers that evidence supports a specific descriptive, interpretive, or causal claim.
- Data access, which affords readers access to the evidence or data used to support empirical research claims.

Enabling reuse involves both technical actions (i.e., converting proprietary or rare file formats to common, open standard file formats) as well as judgment calls that are specific to the context of the project. For example, selecting file formats recommended by the Library of Congress (https://www.loc.gov/preservation/resources/rfs/) can enable both access in the future and access by those using different software and hardware.

Benefits and challenges of sharing qualitative data

While qualitative data can help explain phenomena and describe the world, it doesn't have the same history of preservation and sharing that quantitative data does (Qualitative Data Repository [QDR], n.d.). Because of this, researchers may not consider depositing their data once their research projects are completed, and if they do, researchers and curators may have difficulties preparing the data for deposit.

Similar to the sharing of quantitative data, sharing qualitative data provides numerous benefits for researchers. As will be discussed later on in this primer, the increased trust and transparency and ability to reuse the data are among the key benefits to sharing qualitative data. Encouraging researchers to deposit their files could aid in the preservation of the data, rather than having them be disposed of once the particular research project has been completed (QDR, n.d.). Preserving the qualitative data could allow for easier access to data for longitudinal studies by multiple research teams that currently may be difficult to complete. In addition, sharing qualitative data can allow for a deeper analysis with the ability to have more individuals study the data and provide additional insights (Elman and Kapiszewski, 2017). Finally, as long as deposited data is publically available, it could provide students and researchers learning how to analyze qualitative data a useful dataset to practice on and examine (Elman and Kapiszewski, 2017).

Although there are benefits to sharing qualitative data through depositing it, there are also challenges. The primary one is that in the United States, data generated in the course of qualitative research is not broadly shared or deposited in repositories and researchers may not think about that possibility (QDR, n.d.). Numerous rounds of outreach to researchers conducting qualitative research may be necessary. This could range from letting them know that your repository accepts qualitative data to working with them to prepare their IRB submissions. If researchers express interest in depositing their data without planning for it when going through the ethics review process, they may run into difficulties.

Reproducibility, transparency, and maximizing reuse

Criticisms of qualitative research have generally misunderstood the fundamental differences in philosophy, paradigm, and methods between quantitative and qualitative research. Creswell (2007) proposes that the ultimate goal of qualitative research is understanding (rather than producing generalizable knowledge). The ways in which the validity of qualitative research is evaluated depends on the role and perspective of the reader, participant, researcher, and other stakeholders. Rather than being judged by generalizable criteria, qualitative research is conducted with the assumption that the findings will be valid in some cases and less so in others. According to Lincoln & Guba (1985), the terminology used to describe validity for quantitative research should not be applied to the naturalistic approach taken in qualitative studies. Instead, they propose the use of terms such as credibility, authenticity, transferability, dependability, and confirmability. A more recent list comes from a synthesis of validation approaches by Whittemore, Chase, and Mandle (2001), who found four primary criteria: credibility, criticality, authenticity, and integrity. Creswell (2007) describes eight strategies for supporting validation as a process: prolonged engagement and persistent observation in the field; triangulation (of data sources, methods, investigators, and theories); peer review or debriefing for an external check; negative case analysis; clarifying researcher bias from project inception; member checking; rich, thick description to allow readers to make decisions about transferability; and external audits. He recommends that qualitative researchers use at least two of these strategies in any given study.

The concept of reproducibility is based on a positivistic approach to research. It is defined as the ability to produce the same results when given the same data and methods. Replicability is a related concept which focuses on obtaining the same results when applying the same methods to a different sample or dataset. In contrast to quantitative research, rigor in qualitative research is based on transparency, credibility, reliability, comparability, and reflexivity (Saumure & Given, 2008). The validity, or credibility, of a qualitative study depends upon the selection of methods suitable for the research question, rigorous methods for gathering and analyzing multiple sources of high-quality data, and the researchers self-awareness of their assumptions, biases, and influence upon the study (Patton, 1999). Reliability in a study suggests that similar results would be obtained using similar participants and research methods. For a study to be comparable, researchers need to ensure that all voices in that study are represented. Reflexivity describes the work of the researcher to identify and report how they may have influenced the results. Transparency is an overarching issue in qualitative research. Saumure & Given (2008) describe it as “clarity in describing the research process”, while Hiles (2008) takes a more expansive view. However, both emphasize the importance of transparency with respect to the process, rather than the findings. Transparency requires researchers to provide a comprehensive description of their process, or an audit trail, that allows for evaluation of the suitability of the method for the research question and replication by others.

Characteristics of qualitative research include a natural setting, the researcher as a key instrument, multiple sources of data, inductive data analysis, consideration of participant meaning, emergent design, a theoretical lens, use of interpretive inquiry, and a holistic account. In particular, curators should keep in mind that the design of qualitative research is emergent, rather than strictly defined to test *a priori* hypotheses. As such, documentation of the design is crucial for those who want to evaluate or extend the research, or reuse the data.

Why readme files are important for qualitative data

A well-documented project is crucial for ensuring transparency and reproducibility in qualitative research. Replication data, which includes the raw data, codebooks, and other components required for the original analysis, allow other researchers to replicate and confirm the results of the study.

Research transparency has three dimensions: data, analytic, and production transparency (Moravcsik 2014). In short:

- Data transparency ensures the access of data to other researchers as appropriate.
- Analytic transparency requires providing clear guidelines on how the data were analyzed.
- Production transparency necessitates access to the methods by which particular bodies of cited evidence, arguments, and methods were selected.

When describing the data, analysis, and project, it is important to have robust description that covers the following:[2]

- Data level transparency descriptive information should include:
    - Metadata schema applicable/used in this dataset
    - Parameters and/or variables used
    - Column headings for tabular data
    - Codes or symbols used to record missing data
    - Other specialized formats, abbreviations, or symbols used.
    - Retention information: How long should the data be preserved? Is there a funder requirement?
    - Instruments used in collecting or analyzing data. Include software version information when applicable.
- Analytic level transparency descriptive information should include:
    - Research design and methodology
    - Codebook and other analytic tools
    - Particular information about software utilized, including which version(s).
    - Theoretical framework
    - Levels of coding for analysis
- Production level transparency descriptive information should include:
    - Project history, aims, objective, and hypotheses
    - Structure of data files and relationships between
    - Data confidentiality, access, and licenses
    - Related publications, presentations, or other research outputs
    - Any modifications made over time
- Other study level descriptive information might include:
    - Project name, funding agency, grant award number(s).
    - Every investigator's name, institutional affiliation, role, and ID (e.g., ORCID)
        - Be clear on how the study operated: what were the distinct roles and responsibilities? Is there a distinction between research and authorship in the citation or acknowledgements?
    - Right information (i.e., appropriate license information)
- Potential Metadata standards to consider:
    - DDI -- The Data Documentation Initiative
- Sample template readme files
    - https://data.research.cornell.edu/content/readme
    - https://dataworks.iupui.edu/themes/DataWorks/txt/IUPUI-DataWorks_ReadmeTemplate.txt

Example qualitative datasets and sample citations

- Data for: Exploring sources of insecurity for Ethiopian Oromo and Somali women who have given birth in Kakuma Refugee Camp: A qualitative study (https://data.qdr.syr.edu/dataset.xhtml?persistentId=doi:10.5064/F62T7NYQ)

[2] This information can be stored in a readme or a related publication; however, ensure there are sufficient connections between the two.

  - Sample citation: Lalla, Amber. 2020. "Data for: Exploring sources of insecurity for Ethiopian Oromo and Somali women who have given birth in Kakuma Refugee Camp: A qualitative study". Qualitative Data Repository. https://doi.org/10.5064/F62T7NYQ. QDR Main Collection. V1. UNF:6:DkftR3RiyRnPMueLJQX1jg== [fileUNF]
- Magdalene Oral History Project (https://repository.dri.ie/catalog/dn39x152w)
  - Sample citation: O'Donnell, Katherine, Pembroke, Sinead, & McGettrick, Claire. (2015) Magdalene Oral History collection, Digital Repository of Ireland [Distributor], Irish Qualitative Data Archive [Depositing Institution], https://doi.org/10.7486/DRI.dn39x152w
- World Within Walls collection (https://repository.dri.ie/catalog/5999vb192)
  - Sample citation: Health Service Executive. (2015) World Within Walls collection, Digital Repository of Ireland [Distributor], Irish Qualitative Data Archive [Depositing Institution], https://doi.org/10.7486/DRI.5999vb192
- Data for: When do the dispossessed protest? Informal leadership and mobilization in Syrian refugee camps (https://doi.org/10.5064/F6CN723S)
  - Sample citation: Clarke, Killian B. 2018. "Data for: When do the dispossessed protest? Informal leadership and mobilization in Syrian refugee camps". Qualitative Data Repository. https://doi.org/10.5064/F6CN723S. QDR Main Collection.
- Data for: Authoritarian apprehensions: Ideology, judgment, and mourning in Syria (https://doi.org/10.5064/F63776W4)
  - Sample citation: Wedeen, Lisa. 2019. "Data for: Authoritarian apprehensions: Ideology, judgment, and mourning in Syria". Qualitative Data Repository. https://doi.org/10.5064/F63776W4. QDR Main Collection. V1. UNF:6:xRWn9II/O2ynGGA55KH+OQ== [fileUNF]
- Fifty Victorian Era Novelists Authorship Attribution Data (http://dx.doi.org/10.7912/D2N65J)
  - Sample citation: Gungor, A. (2018). Fifty Victorian Era Novelists Authorship Attribution Data. IUPUI University Library. http://dx.doi.org/10.7912/D2N65J

Workflow based on the Data Curation Network CURATED steps

Use this guide along with other primers, such as the Human Subjects Data Essentials Primer and Curation of Data Collected via Informed Consent, to determine the best next steps for your repository and institution. You can ensure you have completed a job-well-done with the Data Curation Network's CURATE checklist:

C- Check

- ❏ Completion of Consent Form Assessment
- ❏ Screen files to ensure that no sensitive data is included (See Human Subjects Data Essentials Primer)
- ❏ Check for the following necessary files (and link to above list of recommended files)
  - ❏ Original source data, when possible, and as long as it is de-identified
  - ❏ Processed data (e.g., transcripts), if applicable
  - ❏ Codebook
  - ❏ Readme
  - ❏ Instrument for data collection (e.g., survey, interview, etc), if applicable
  - ❏ Documentation generated in the Qualitative Data Analysis Software (QDAS) application (e.g., Nvivo, atlas.TI, etc.): memos, notes, networks, classifications

U- Understand

- ❏ Understand the type(s) of data submitted and whether they were analyzed using qualitative methods or tools (QDAS).
    - ❏ Try opening the QDAS files if possible.
- ❏ Refer to the documentation (i.e., codebook, hierarchy of nodes) for the coding or tagging schema
- ❏ Consult the methods as described in the funding proposal, manuscript, or other documentation to understand how the data were generated, coded, and analysed.

R-Request

- ❏ Request missing information or changes
    - ❏ Consult with the depositor to discuss what minimal documentation is necessary for others to evaluate and reuse the dataset (see the Curation Checklist section above)

A-Augment

- ❏ Augment metadata for findability
    - ❏ Work with the depositor to create a readme file that contains project information, a file directory, information about data sources
- ❏ Augment the internal record with a copy of the consent form or agreement and potentially note in the readme file that the consent indicated public sharing of the data. (See Human Subjects Data Essentials Primer)
- ❏ If the readme was created by the depositor, check for the elements listed above (See “Why readme files are important for qualitative data”)

T-Transform

- ❏ Transform file formats for reuse
    - ❏ Many QDAS are proprietary, so interoperability between them is not expected. See the software-specific primers (i.e., Atlas.ti, NVivo) for guidance in extracting key information from QDAS.
    - ❏ When the original source files are not available or are in proprietary formats, consider exporting them to common, openly defined file formats.
    - ❏ When possible, convert files into open-definition, common file formats based on the recommendations from the Library of Congress: https://www.loc.gov/preservation/resources/rfs/RFS%202019-2020.pdf.

E-Evaluate

- ❏ Evaluate for FAIRness (See evaluation tools at https://fairshake.cloud/rubric/8/assessments/ or https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/#!/)
    - ❏ Evaluate the record to ensure the data is Findable, Accessible, Interoperable, and Reusable. For qualitative data, the focus should be on exporting as much contextual information about how the inquiry process was conducted and using accessible file formats.
        - ❏ Findable - Ensure the dataset is described with relevant metadata, including an abstract; assign a DOI; choose an appropriate repository to make the data available.
        - ❏ Accessible - This varies depending on the file formats; refer back to transform point about using common, openly defined file formats.
        - ❏ Interoperable - Source files should be in common, openly defined file formats, but project files will likely be proprietary; link to related publication(s) when possible; use a formal, accessible, and applicable vocabulary to describe the dataset.

- ❑ Reusable - Highly dependent on the documentation; source data may be reusable even if the process cannot entirely be replicated; a clear and accessible data usage license is specified in the metadata and readme; documentation should include provenance of original source data.

D-Document

- ❑ Document (the curation process)
  - ❑ The files received - the original data and any documentation provided
  - ❑ Transformations - what files were transformed
  - ❑ Provenance of original source data
  - ❑ Embargo protocols, if applicable
  - ❑ Controlled sharing protocols, if applicable
  - ❑ Correspondence with depositor (may need to define the scope; limit to those contained within specific systems?)
  - ❑ Evaluation of FAIRness (find link to the rubric, if it exists)
  - ❑ Repository record information (i.e., handle/PURL, metadata, etc.)
  - ❑ Any additional requirements at repository in question

Bibliography

About the Qualitative Data Repository | Qualitative Data Repository. (n.d.). Retrieved March 24, 2020, from https://qdr.syr.edu/about

Corti, L. (2018, July 5). *Show me the data: Research reproducibility for qualitative methods*. NCRM Research Methods Festival, University of Bath.

Creswell, J. W. (2007). *Qualitative inquiry & research design: Choosing among five approaches (2nd Edition)*. Sage Publications.

Elman, C., & Kapiszewski, D. (2017, August 9). Benefits and Challenges of Making Qualitative Research More Transparent. Inside Higher Ed. https://www.insidehighered.com/blogs/rethinking-research/benefits-and-challenges-making-qualitative-research-more-transparent

Flick, U. (2014). Mapping the field. In Flick, U. *The SAGE handbook of qualitative data analysis* (pp. 3-18). London: SAGE Publications Ltd doi: 10.4135/9781446282243

Flick, U. (2018). Doing qualitative data collection – charting the routes. In Flick, U. *The sage handbook of qualitative data collection* (pp. 3-16). London: SAGE Publications Ltd doi: 10.4135/9781526416070

Hiles, D. R. (2008). Transparency. In *The Sage encyclopedia of qualitative research methods*. Sage. https://dx-doi-org.proxy.ulib.uits.iu.edu/10.4135/9781412963909.n467

Lincoln, Y. S. & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

Moravcsik, A. (2014). Transparency: The Revolution in Qualitative Research. *PS: Political Science & Politics*, *47*(1), 48–53. https://doi.org/10.1017/S1049096513001789

OpenAIRE: How to select a data repository? https://www.openaire.eu/opendatapilot-repository-guide

Patton, M. Q. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Services Research*, *34*(5 Pt 2), 1189–1208.

Saumure, K. & Given. L. M. (2008). Rigor in Qualitative Research. In *The Sage encyclopedia of qualitative research methods*. Sage. https://dx-doi-org.proxy.ulib.uits.iu.edu/10.4135/9781412963909.n409