QDS
Qualitative Data Sharing
PROJECT

Qualitative Data Sharing: A Roadmap and Resources to Facilitate Responsible and Ethical Data Sharing

ICPSR

(I²)
Institute *for* Informatics

BRC
BIOETHICS RESEARCH CENTER
@
Washington University in St. Louis

Introductions

Jessica Mozersky, PhD
Co-Investigator
Bioethics Research Center

Albert Lai, PhD
Co-Investigator
Institute for Informatics (I^2)

Sara Britt
Data Project Manager
ICPSR

James DuBois, DSc, PhD
Principal Investigator, BRC

Aditi Gupta, PhD
Co-Investigator, I^2

Amy Pienta, PhD
Co-Investigator, ICPSR

Washington University in St. Louis
School of Medicine

QDS
Qualitative Data Sharing
PROJECT

Relevant Disclosures

- The Wash U Qualitative Data Sharing (QuaDS) Software is now called **De-ID** with exclusive licensing rights held by SCRC's The Institute for Mixed Methods Research (the publisher of Dedoose).

- Drs. DuBois, Lai, Gupta and Mozersky will receive part of the profits from any sales of this software.

- Sara Britt works for a non-profit data repository that charges fees for some data deposits.

- All work described in this grant was supported by the National Institutes of Health (NHGRI, R01HG009351, PI DuBois)

Outline

- About the QDS project
- Empirical Data findings
- Guidance based on our experience and lessons learned
- QUADS Software development and demonstration
- ICPSR Repository experience

QDS
Qualitative Data Sharing
PROJECT

About the QDS Project

QDS
Qualitative Data Sharing
PROJECT

About the QDS Project

Sharing Qualitative Research Data: Identifying and Addressing Ethical and Practical Barriers *(NHGRI, R01HG009351)*

- **Qualitative methods** (i.e., interviews, focus groups) enable unique insights into **health related** behaviors, emotions, and attitudes including highly sensitive issues
- **Revisions to regulations require data to be shared at the time of publication**
  - NIH Data Sharing Policy effective January 2023 (NOT-OD-21-013)
  - White House OSTP Data Sharing Policy effective December 2025
  - National Science Foundation

Our project is …

- US based (populations, regulatory frameworks)
- Focused on **health** research and data collection
- Focused on interviews and focus groups
    - Not video or ethnographic field notes currently

Benefits of QDS

- Enable secondary analyses and other syntheses
- Cost effective
- Potentially prevent over burdening participants
- Training with real world data
- Foster public trust through transparency
  - Not replication

About the QDS Project

- **Aim 1:** Stakeholder engagement to identify ethical and practical barriers and facilitators of QDS
  - Qualitative Researchers, IRB members, Data repository curators, Research participants
- **Aim 2:** Qualitative data sharing trial with formative evaluation of QuaDS software
  - 30 QRs de-identified and deposited datasets
- **Aim 3:** Development, evaluation, and dissemination of a QDS Toolkit containing stakeholder-informed guidelines and QuaDS software

Aim 1 Qualitative Findings (N=120)

- IRB members (N=30)
  - Supportive so long as data adequately de-identified and consent obtained
- Data repository curators (international) (N=30)
  - Unequivocally supportive but felt unprepared
- Qualitative Researchers (N=30)
  - Concerned, ambivalence, lack knowledge
- Research participants (N=30)
  - Supportive if de-identified and shared with other researchers
  - Expectation that data already shared

Quantitative Survey Findings (N=425 Qualitative Researchers in the US)

- Nearly evenly split between support and opposition
  - Oppose (41%) vs. support (49%) sharing
- **96%** had never shared QR data with a repository
- Biggest concerns:
  - lack of permission (87%)
  - data sensitivity (85%)
  - participant trust (82%)
  - IRBs (79%)
  - data de-identification concerns (79%)
- Health science researchers more supportive than researchers from anthropology, sociology, other fields

QDS
Qualitative Data Sharing
PROJECT

QDS Planning Guide

Key Considerations for Project Planning & Management

&

Lessons Learned from QDS Project

QDS in a Project Timeline

	PRE-AWARD	STUDY SETUP	DATA COLLECTION	DATA ANALYSIS	REPORTING & PUBLICATION	
1: Clarify what "Data Sharing" means for your project	🟩					
2: Include Data Sharing in Funding Proposal	🟦					
3: Find a Location for Shared Data	🟩	🟩				
4: Include Plans to Share in IRB		🟦				
5: Execute Data Sharing Agreements		🟩	🟩	🟩		
6: Account for Data Sharing in Project Management		🟦	🟦	🟦		
7: De-Identify Data			🟩	🟩		
8: Deposit Data					🟦	
9: Track Data Sharing Impact						🟩

QDS
Qualitative Data Sharing
PROJECT

QDS Pilot Project

DATA PREPARATION

- Execute Agreements
- Upload identifiable data to secure WUSTL server

DATA DE-IDENTIFICATION

- QauDS Software flags indirect & direct identifiers
- Review de-identified files

DATA DEPOSIT

- Execute Restricted-Use Data Deposit & Dissemination Agreement (RUDDDA)
- Deposit the final de-identified data files in ICPSR
- ICPSR archives data and applies access restrictions

QDS
Qualitative Data Sharing
PROJECT

QDS Pilot Project

Dataset Topics:

- Sex, Relationships, & Reproductive Health
- Mental Health & Substance Use
- Violence & Abuse
- Cancer survivorship

Datasets Enrolled in QDS Pilot Project (n=28)

Data Type*		
Interviews	21	75%
Focus Groups	12	43%
Funding Source		
NIH	13	46%
No External Funding	11	39%
Other	4	14%
Sample Population*		
Healthy Individuals	15	54%
Patients	6	21%
Children	2	7%
Older Adults (65+)	3	11%
Individuals with Stigmatized Diagnoses	5	18%
Economically Disadvantaged Individuals	7	25%
Other	17	61%

***** Survey item enabled check all that apply.***

QDS
Qualitative Data Sharing
PROJECT

STEP 1: Include Data Management & Sharing in Funding Proposal

- **Consider at the outset of a proposal what your plans are and look for existing guidance**
- Data Sharing Plan
- Data Management Plan

STEP 1: 3 Key elements of planning

Data sharing may impact other sections of the funding proposal:

- **Budget & Milestones**
  - Funder may allow budget
  - NIH requires costs incurred during project period
- **Consent Forms (more on this later)**
  - Must address data sharing
    - Include relevant details (restricted vs open access, where shared, who can access)
- **Data management and documentation**
  - Follow best practices

QDS
Qualitative Data Sharing
PROJECT

STEP 2: Which materials do I share?

! Always check with your funding agency and data repository for updated guidance or expectations regarding data sharing.

Files to SHARE:

- **De-Identified Data Files**
  - Transcripts & accompanying data
- **Metadata / Supporting Documentation**
  - Study description
  - Data collection & analysis instruments
  - De-identification protocols
  - Bibliography of dataset's publications
  - Contact info original investigator

Files to ARCHIVE:

- **Identifiable Data Files**
  - Audio or video files, raw transcript files
- **Code Applications**

STEP 3: Find a Location for Shared Data

Key Considerations:

- **Cost**
- **Secondary User Access**
    - Public vs Restricted-Use
    - Data Sharing Agreements
- **File Types**

Recommended Locations:

- **Qualitative Data Repositories**
    - Inter-university Consortium for Political and Social Research **(ICPSR)**
    - Qualitative Data Repository **(QDR)**
- **Institutional or Journal Archives**

Benefits of using established repository

- Unique persistent DOI
- Curation and support
- Clear processes for accessing data
- Guidelines for appropriate use by secondary users
- State of the art security
- **Restricted Access**
  - Secondary users must apply to use data
  - Describe data protection plans
  - IRB approved protocol
- Enables higher level of protection
- Requires extra layer of agreements

STEP 4: Permissions to Share, IRB

Check with the IRB or repositories to see if they have guidance or templates. Data sharing plans will impact:

- **Informed Consent**
  - Must be transparent and include relevant details such as how secondary users may access data
  - Should not promise data destruction or never sharing outside original team

- Repositories and IRBs have template language

I didn’t get consent for data sharing … now what?

- Silent consent forms vs explicit prohibition
- Explicit prohibition (e.g. “your data will not be shared outside the original team” “all data will be destroyed”)
  - Data can not be shared without reconsent

Example of Reconsent

- Example from our project
  - Consent prohibited data sharing of highly sensitive data
  - QR recontacted 34 / 39 original participants
  - 19 agreed to data sharing (2 requested to review transcripts)
  - Not all data could be shared, but a portion of it was with participant permission

Ethical vs Regulatory Definitions

- The “common rule”—or federal regulations 45CFR46 on the Protection of Human Subjects—does not consider research with de-identified data to be human subjects research (45CFR46.102(e))
- Removing HIPAA safe harbors meets regulatory requirement and would not be considered human subjects data
- Ethical vs regulatory definition

QDS
Qualitative Data Sharing
PROJECT

STEP 5: Permissions to Share, Institutional Agreements

Begin executing data sharing agreements with repository ***as early as possible:***

- **Use agreement templates provided by repository**
- **Identify & contact the appropriate Institutional Official**
  - Research Contracts Office (not the IRB in most cases)

Your Institutional Official will:

- **Determine data ownership & sharing permissions**
- **Execute data sharing agreements with repository**

! Do not complete the data deposit until agreements are fully executed.

Step 6:
Deidentify Data

What counts as de-identified qualitative data?

- No regulatory standards exist specifically for qualitative data
  - Common Rule provides no specific guidance
- HIPAA Privacy Rule "safe harbor" standard (not research specific)
  - 18 HIPAA identifiers removed (names, addresses, email, phone, device numbers, photographic images, social security number…)
  - Expert determination alternative
- Designed for quantitative data
  - Majority of HIPAA not present in qualitative data

What counts as de-identified qualitative data?

- **Our approach:** remove all 18 HIPAA identifiers & additional variables likely to be found in qualitative data
  - QUADS software flags these

- All variables flagged by QuaDS unlikely to require removal or masking
  - Project specific and require researcher input

"As much as necessary as little as possible"

- Reasonable risk of re-identification = remove
- Leave as many details as possible (while achieving de-identification) to preserve context
- When determining how much to remove, consider consequences:
  - Are the data particularly sensitive?
  - Would re-identification cause social, emotional, legal, or financial harm?
- If yes, err on the side of restricting access

STEP 6: De-Identify Data

Develop a codebook of identifiable information, and keep it updated throughout data collection:

- **Determine what is identifiable in the context of the study**
  - As much as necessary and as little as possible
- **Follow rules & regulations**
  - 18 HIPAA Safe Harbor identifiers, funding agency guidance
- **Use consistent & trackable replacements**
  - "I got my master's at **WashU**. Now I work at **Johns Hopkins**"
  - "I got my master's at **[Mid West University 1]**. Now I work at **[North East University 1]**"

QDS
Qualitative Data Sharing
PROJECT

STEP 6: De-Identify Data

- **Establish a Data Review Process for the Team**
  - Resolve questions, revise codebook, and make ad hoc decisions
- **Be aware of information disclosed in Linked Data Sources**
  - Survey data, publications about the data, or within the same transcript
- **Use De-Identification Support Tools & Services**
  - QuaDS Software
  - Transcription & Repository Services

QDS
Qualitative Data Sharing
PROJECT

Additional Advice …

Clean while you cook!

- Don’t wait until the time of publication to start preparing a data deposit. Reserve time, effort, and resources for data sharing.

Protocolize the data sharing process

- **File naming & storage system**
- **De-identification procedure**
- **Supporting documentation**

- **Your future self will be grateful!**

STEP 7: Deposit Data

- **Upload & submit deposit to repository**
  - Data files & supporting documentation
- **Receive dataset citation**
- **Stay in contact with repository throughout curation process**
  - Disclosure Risk Review

If applicable, update data deposit throughout the project as new data becomes available.

STEP 8: Track Data Sharing Impact

- **Cite dataset in subsequent publications & annual reports**

- **Check user metrics**
  - Views, downloads, etc.

Future Considerations

- There is no accepted standard for when qualitative data are adequately de-identified

- Risk of Participant Re-identification is real
    - Open access may present a real challenge

- Mixed methods data may present greater challenges

Future Considerations

- Our experience is US based
- Goal is transparency not reproducibility
- Not appropriate in all contexts
  - Some data can not be shared safely and responsibly
  - In rare cases, withholding a particular transcript may be necessary

Future Considerations

- Risk that deposited data are not useful is real
  - Data quality is necessary for secondary use

- Secondary usability: Tradeoffs may be needed to ensure data are useful to secondary users
  - Overly de-identified data may not be usable
  - Metadata and documentation
  - Restricted access may enable less stripping of identifiers
  - Our review does not suggest deep relationships in the health sciences

Future Considerations

- Broad Community and stakeholder input will be needed
  - Control regarding how data are used may be lessened
    - CARE principles for indigenous data governance (**C**ollective benefit, **A**uthority to control, **R**esponsibility, and **E**thics)
- Historical Mistrust

QDS Project Resources

ICPSR Series Page for the QDS Project:

https://www.icpsr.umich.edu/web/ICPSR/series/1780

Qualitative Data Repository:

https://qdr.syr.edu

PI: Jim DuBois (after project funding ends)

duboisj@wustl.edu

QDS
Qualitative Data Sharing
PROJECT

Overview of QuaDS Software and Demo

QDS
Qualitative Data Sharing
PROJECT

About the QuaDS Software

- Developed a support software to help qualitative researchers anonymize and validate their datasets
- Approved for HIPAA/PHI storage
- Example qualitative transcripts: narratives, interviews, field notes, or focus group data
- User interface and backend algorithm developed by Washington University School of Medicine software development team

QuaDS Software Features

- QuaDS software allows users to review, validate and edit the identifiers flagged by the software before downloading the final de-identified transcript. Some of the key features are:
    - Automatically highlight text that are believed to be identifiers based on syntactic or grammatical features
    - Replace identifiers with contextual substitution texts
    - For each identified changed, we track the text and number of replacements, thus producing an anonymization log

U.S. De-Identification Standards

U.S. regulatory standards

- HIPAA “18 safe harbor identifiers” or “expert determination” (HSH)
  - Name, address, MRN, SSN, etc.

- May not suffice for qualitative data:
    - E.g., “CEO of Purina since 2018” or only “male, Hispanic, psych nurse in St. Louis”

HIPAA “18 safe harbor identifiers”

1. Names
2. All geographic subdivisions smaller than a state
3. All elements of dates (except year) for dates
4. Telephone numbers
5. Vehicle identifiers and serial numbers, including license plate numbers
6. Fax numbers
7. Device identifiers and serial numbers
8. Email addresses
9. Web Universal Resource Locators (URLs)
10. Social security numbers
11. Internet Protocol (IP) addresses
12. Medical record numbers
13. Biometric identifiers, including finger and voice prints
14. Health plan beneficiary numbers
15. Full-face photographs and any comparable images
16. Account numbers
17. Any other unique identifying number, characteristic, or code,
18. Certificate/license numbers

Identifier Category	Definition
HIPAA Safe Harbor (HSH)	All 18 HIPAA defined categories[40]
NON-HIPAA	
Location	Geographic area larger than a state including country: “I was born on the East Coast”.
Age/Dates	Age in years, months, or weeks: “The baby was four weeks old”. Includes non-HSH dates or holidays such as “Thanksgiving” or “Christmas Day”.
Numbers	All non-HSH numbers: “He weighed over 600 pounds” or “She had thirteen children” (except “one”)
Organization/ Institution	Names of institutions or organizations such as “Barnes-Jewish Hospital” which is not an address and constitutes multiple locations, and proper names: “Pfizer”, “Red Cross”.
Race, Ethnicity, Nationality, Indigenous Status	NIH racial/ethnic categories, indigenous status, nationality: “My patient was from Haiti”.
Sexual Orientation / LGBTQI	Sex or gender that is not cis (including intersex, transsexual, asexual) and non-heterosexual orientation: “We got married the day gay marriage ban was lifted”.
Rare Diseases	Rare diseases
Other	Rare events and other rare references not captured under any existing category and that are unlikely to be captured by automation such as “He won the Olympic gold medal for swimming in Houston” or “Nobel laureate in 1995”

QuaDS Backend

DE-IDENTIFICATION PIPELINE DEVELOPMENT AND EVALUATION

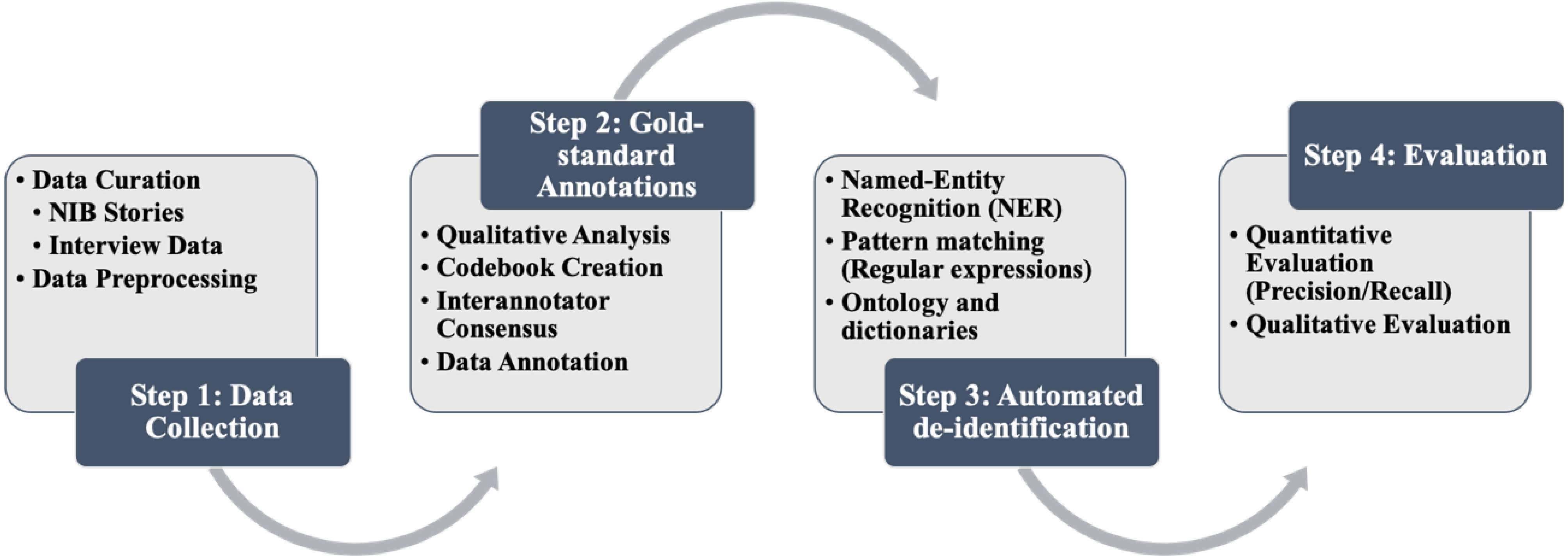
De-identification Pipeline Development

- We developed and validated a pipeline for deidentifying qualitative research data using automated computational techniques.

- An in-depth analysis and qualitative review of different types of qualitative health research data were conducted.

- Utilized natural language processing (NLP) methods including named-entity recognition, pattern matching, dictionary, and regular expression methods to deidentify qualitative texts.

De-identification Model Development



Step 1: Data Collection

- Data Curation
  - NIB Stories
  - Interview Data
- Data Preprocessing

Step 2: Gold-standard Annotations

- Qualitative Analysis
- Codebook Creation
- Interannotator Consensus
- Data Annotation

Step 3: Automated de-identification

- Named-Entity Recognition (NER)
- Pattern matching (Regular expressions)
- Ontology and dictionaries

Step 4: Evaluation

- Quantitative Evaluation (Precision/Recall)
- Qualitative Evaluation

Descriptive statistics of the two datasets used in the study and the number (%) of identifiers (HSH and non-HSH) extracted using the NLP pipeline from each set; and gold-standard evaluation of the NLP system.

	Dataset Name (Number of Files)	Number of Tokens	Number of Identifier Tokens (%)	Precision	Recall	F1 Score
Training/ Learning Datasets	NIB Stories (6 files)	12620	389 (3 %)	0.93	0.92	0.93
	QDS Interviews (9 files)	85590	650 (1 %)	0.98	0.90	0.94
Testing Datasets	NIB Stories (25 files)	48807	858 (2%)	0.93	0.98	0.95
	QDS Interviews (30 files)	139323	998 (1%)	0.97	0.95	0.96
Total	70	286340	2888 (1%)	0.95	0.96	0.96

Enabling qualitative research data sharing using a natural language processing pipeline for deidentification: moving beyond HIPAA Safe Harbor identifiers. Gupta A, Lai A, Mozersky J, Ma X, Walsh H, DuBois JM. JAMIA Open. 2021

Results and Conclusion

- The majority of identifiers in qualitative data are non-HSH and not captured by existing systems.
- The NLP deidentification pipeline had a consistent F1-score of ~0.95 for both datasets.
- The results of this study demonstrate that NLP methods can be used to identify both HSH identifiers and non-HSH identifiers.

QuaDS User Interface

CLOUD-BASED USER INTERFACE AND DEMO

Software Details

- Cloud based web application hosted on Microsoft Azure platform
- Accessible to approved researchers and study participants
- Support TXT file format
- Stay tuned for next steps: https://qdstoolkit.org/quads/

3-Step Process

1. Upload text files for de-identification
2. Review and approve the de-identification tags
3. Download anonymized files

QDS
Qualitative Data Sharing
PROJECT

Demo

Software User Evaluation

- 30 researchers across US enrolled in our research study to use and evaluate the QuaDs software and de-identify their datasets

- Data collected:
  - User Survey
  - Log Analysis
  - De-identified text files

QuaDS Survey Analysis Results

- We iteratively improved the software based on the feedback received from the qualitative researchers.
- Summary feedback:
  - Most of the participants showed high confidence in QuaDs to maintain privacy of their participants.
  - Most participants were satisfied with the experience of using the QuaDS Software and said they will be interested in using it in future for their research work.

QDS Project Resources

ICPSR Series Page for the QDS Project:

https://www.icpsr.umich.edu/web/ICPSR/series/1780

Qualitative Data Repository:

https://qdr.syr.edu

PI: Jim DuBois (after project funding ends)

duboisj@wustl.edu

De-ID Software

The WashU Qualitative Data Sharing Software (QuaDS) is now called **De-ID**

Exclusive licensing rights held by SCRC's Institute for Mixed Methods Research. (The Publisher of Dedoose)

For more information on its development and availability, please email contact@immrglobal.org

For more information on the project overall, contact the principal investigator, Jim DuBois at **duboisjm@wustl.edu**

QDS
Qualitative Data Sharing
PROJECT

Questions?

BRC
BIOETHICS RESEARCH CENTER
@
Washington University in St. Louis

(I^2)
Institute *for* Informatics

M | ICPSR

CONTACT US

QDS@WUSTL.EDU

Qualitative Data Sharing

Concerns with sharing qualitative data

- sensitivity of data
- participant confidentiality
- data ownership
- time burden and cost of de-identifying data

Requirements for data sharing

- improve transparency
- promote secondary data analysis
- facilitate research training

Best Practices

- Step 1: Review and Process Data Deposit Agreement

Best Practices

- Step 2: Upload De-identified Data to ICPSR

Best Practices

- Instructions for Depositing Data at ICPSR

Deposit Manager

Workspace / ICPSR-186701

Sharing Qualitative Research Data: Interviews with Research Participants, 2018 [Deposit Queued]

☐	Name	Type	Size	Last Modified	Actions
☐	De-Identified-Data			5/16/2023 12:19:42 PM	
☐	Supporting-Documenta...			5/16/2023 12:17:31 PM	

Deposit Manager

− Project Description

Principal Investigator(s) (required) ?

James M. DuBois, Washington University in St. Louis. School of Medicine

Summary (required) ?

The Qualitative Data Sharing Project: Aim1A Interviews with Research Participants, 2018 data package provides data from surveys and semi structured, in-depth qualitative interviews with 30 individuals who have experience participating in sensitive qualitative research studies to explore their understanding and concerns about qualitative data sharing. Understanding participants' views about sharing their qualitative data may ensure that federal and institutional data sharing policies, informed consent processes, and data-sharing plans are developed in ways that are consistent with the preferences of research participants and maintain public trust in research.

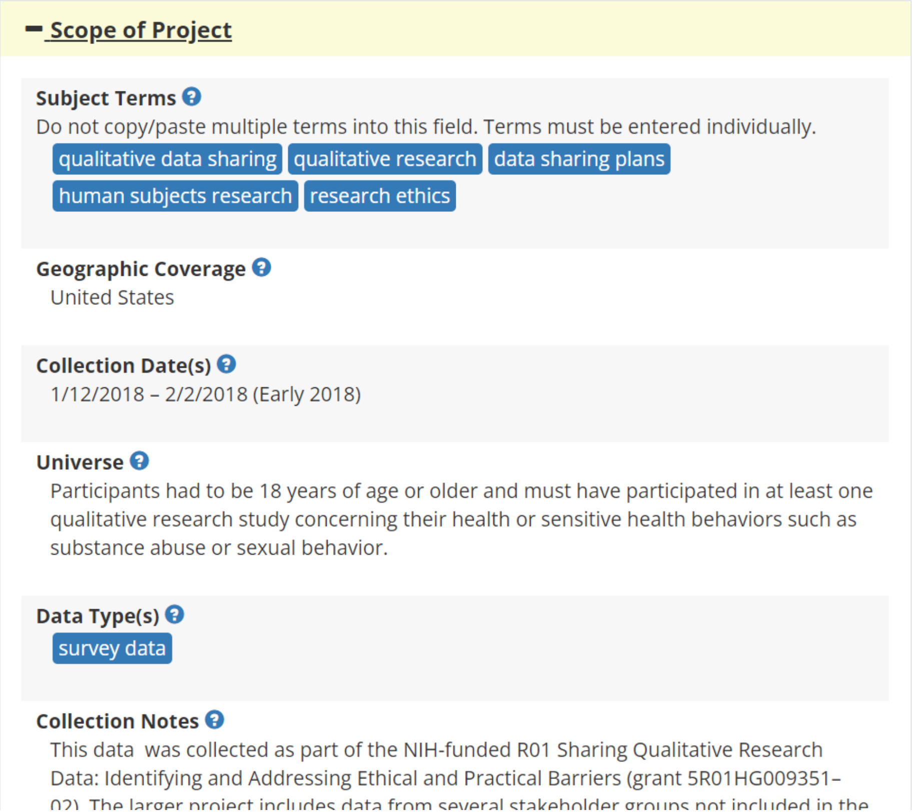
Funding Sources ?

NIH National Human Genome Research Institute (NHGRI) (R01HG009351)

Grant Manager(s) ?

Heidi Walsh, Washington University in St. Louis. School of Medicine

Deposit Manager



Scope of Project

Subject Terms

Do not copy/paste multiple terms into this field. Terms must be entered individually.

Geographic Coverage

United States

Collection Date(s)

1/12/2018 – 2/2/2018 (Early 2018)

Universe

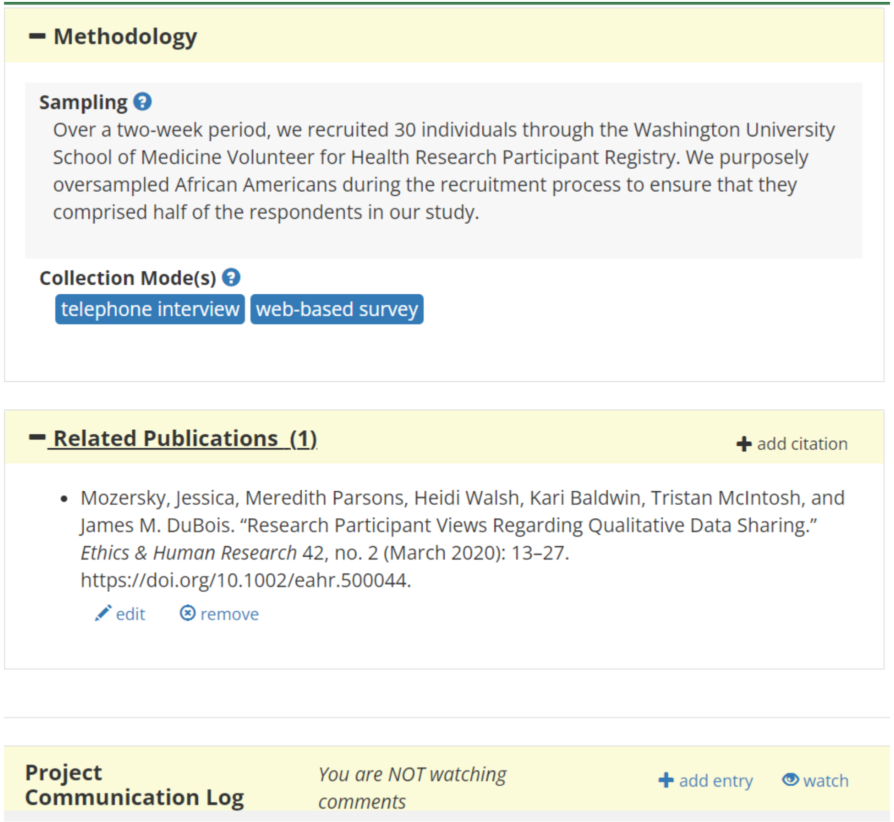
Participants had to be 18 years of age or older and must have participated in at least one qualitative research study concerning their health or sensitive health behaviors such as substance abuse or sexual behavior.

Data Type(s)

Collection Notes

This data was collected as part of the NIH-funded R01 Sharing Qualitative Research Data: Identifying and Addressing Ethical and Practical Barriers (grant 5R01HG009351–02). The larger project includes data from several stakeholder groups not included in the

Deposit Manager



Sampling

Over a two-week period, we recruited 30 individuals through the Washington University School of Medicine Volunteer for Health Research Participant Registry. We purposely oversampled African Americans during the recruitment process to ensure that they comprised half of the respondents in our study.

Collection Mode(s)

telephone interview | web-based survey

- Mozersky, Jessica, Meredith Parsons, Heidi Walsh, Kari Baldwin, Tristan McIntosh, and James M. DuBois. “Research Participant Views Regarding Qualitative Data Sharing.” *Ethics & Human Research* 42, no. 2 (March 2020): 13–27. https://doi.org/10.1002/eahr.500044.

Project Communication Log

You are NOT watching comments

Best Practices

- Acceptable Data and Documentation Formats

Best Practices

- Data disclosure review

Best Practices

- Special Considerations

Best Practices

- Levels of Restriction

Best Practices

- Curation

Qualitative Data Sharing Project Series

- Housed at ICPSR
- Allowed us to pilot the process for using QuaDS and for archiving qualitative data

SCAN ME

ICPSR

Find out more

SCAN ME

Website

qdstoolkit.org

Email

icpsr-help@umich.edu