Knight, Gareth (2018) Data Management for interview and focus group resources in health. London School of Hygiene & Tropical Medicine. (Unpublished)

Downloaded from: http://researchonline.lshtm.ac.uk/4646631/

DOI: 10.17037/PUBS.04646631

Usage Guidelines

Please refer to usage guidelines at http://researchonline.lshtm.ac.uk/policies.html or alternatively contact researchonline@lshtm.ac.uk.

Available under license: http://creativecommons.org/licenses/by/2.5/
DATA MANAGEMENT FOR INTERVIEW AND FOCUS GROUP RESOURCES IN HEALTH

Catalogue Information:

<table>
<thead>
<tr>
<th>Field</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title:</td>
<td>Data Management for interview and focus group resources in health</td>
</tr>
<tr>
<td>Document type:</td>
<td>Guide</td>
</tr>
<tr>
<td>Creator:</td>
<td>Gareth Knight</td>
</tr>
<tr>
<td>Keywords:</td>
<td>research data management, interviews, focus groups, health data, social science</td>
</tr>
<tr>
<td>Description:</td>
<td>This Guide to Good Practice provides advice to LSHTM researchers managing qualitative data acquired through interviews and focus groups. It outlines questions to be considered at each stage, management approaches that may be taken, and resources where further information may be found.</td>
</tr>
<tr>
<td>Language:</td>
<td>English</td>
</tr>
<tr>
<td>Rights:</td>
<td>Creative Commons Attribution 4.0 International License</td>
</tr>
</tbody>
</table>

Version control:

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Change description</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>19 Feb 2018</td>
<td>First version</td>
<td>Gareth Knight</td>
</tr>
<tr>
<td>1.1</td>
<td>12 June 2018</td>
<td>Added extra transcription tool</td>
<td></td>
</tr>
</tbody>
</table>

Feedback
This document will be reviewed and updated on an ongoing basis. To suggest enhancements or amendments contact researchdatamanagement@lshtm.ac.uk.
Contents

Introduction .................................................................................................................. 3

Audio recordings and data protection ................................................................. 3

1. Prepare for data collection ................................................................................. 4
   1.1. Select audio capture device ...................................................................... 4
   1.2. Configure audio settings ......................................................................... 5
   1.3. Ethics and informed consent ..................................................................... 5
   1.4. Data security in the field .......................................................................... 5
   1.5. Reporting lost/stolen devices .................................................................... 7

2. Prepare data for analysis .................................................................................... 8
   2.1. Data security ............................................................................................... 8
   2.2. Transcription ............................................................................................... 8
      2.2.1. Select audio for transcription ............................................................... 8
      2.2.2. Choose transcription convention ....................................................... 8
      2.2.3. Estimate time required ....................................................................... 9
      2.2.4. Assign responsibility: In-house vs. outsource ................................. 9
      2.2.5. Locate software and hardware .............................................................. 10
      2.2.6. Quality Assurance ............................................................................. 10
   2.3. De-identification and redaction ................................................................. 11
      2.3.1. Audio recordings ............................................................................... 11
      2.3.2. Written transcripts ............................................................................ 11
   2.4. File formats for analysis ............................................................................ 12
   2.5. Documentation ............................................................................................ 13

3. Preparing for preservation .................................................................................. 13
   3.1. Preservation formats .................................................................................. 14

4. Prepare data for sharing ...................................................................................... 15
   4.1. Identify data to share .................................................................................. 15
   4.2. Select a digital repository to curate and preserve the data ...................... 16
   4.3. Oversight of the decision-making process for data access .................... 16
   4.4. Choosing an access method ...................................................................... 16
   4.5. Determine permitted and non-permitted uses ......................................... 17
   4.6. Adopt file formats suitable for use ............................................................ 17
   4.7. Provide support documentation ............................................................... 17

Definitions .................................................................................................................. 17
Introduction

Qualitative research often involves the performance of interviews and focus groups, each of which can produce a variety of resources, including audiovisual and text-based data. These resources must be carefully managed to ensure that they are captured, processed, preserved and shared in a form that protects research participants, enables the research objective to be achieved and complies with funding, ethical and legal requirements.

This Guide to Good Practice provides advice to LSHTM researchers managing qualitative data acquired through interviews and focus groups. It outlines questions to be considered at each stage, management approaches that may be taken, and resources where further information may be found.

Audio recordings and data protection

Health researchers have a duty of care to participants involved in the research process and must ensure information held about them is managed in an ethical and legally compliant manner.

Staff and students acting on behalf of the School must comply with UK data protection legislation, even if data collection takes place in countries without equivalent legislation. This legislation, as established in the Data Protection Act (DPA) 1998 and its successor, the EU General Data Protection Regulation (GDPR), which will apply in the UK from May 2018, states that living individuals have rights over how information about them is stored and used, and defines a set of mandatory requirements with which those responsible for its management must comply. In-country legislation on the collection, storage and use of personal data that apply to research participants must also be taken into account when performing research outside the UK and working with international partners. In cases where in-country legislation conflicts or goes beyond UK data protection legislation, the in-country requirements will take priority. Information on the deceased are not protected by data protection legislation, however there may be a common law duty of confidentiality to the estate of a deceased person to protect certain information (e.g. sensitive medical information). Researchers should also be sensitive to the impact that use of information on a deceased person may have upon family members.

The School takes the view that qualitative data collected through interviews and focus group discussion constitute Personal Data under data protection legislation. Each person in the audio recording will present specific vocal characteristics - their speech pattern, accent, vocabulary, and other factors – that may be sufficiently distinct to enable identification, either independently or in combination with other information. Statements made may also identify the speaker or other individuals or groups. Although audio processing techniques such as dynamic pitch shifting, background noise addition, and other techniques exist that can be used to make a voice less recognisable, there remains a possibility that some characteristics may remain or that the vocal processing could be reversed. For this reason, it is impractical to anonymise audio and the transcript should be used as a basis for analysis instead.

1. Prepare for data collection

Data collection should be carefully planned to ensure suitable data can be acquired and used for analysis. Key activities to be performed at this stage include:

- Preparation of research questions or other material needed to guide data collection;
- Selection and setup of data capture tools, e.g. audio recording hardware and software;
- Application of security measures to protect data when working in the field;
- Trialling of data collection in controlled conditions and subsequent configuration;
- Preparation of consent form

1.1. Select audio capture device

Audio may be recorded using several device types:

- **Smartphone and a voice recording app:** Smartphones are the preferred method for audio capture at LSHTM, due to their security features (lock screens, encryption, and biometric features) which can be configured to prevent others accessing data held on the device, and internet connectivity which allows data transfer to a secure server when working in the field. A disadvantage of many devices, however, are the inclusion of low-quality microphones that only capture near-by sound. Researchers wishing to use a smartphone are advised to evaluate the capability of their smartphone and purchase an external mic, if appropriate. Several audio recording apps are suitable for academic research, including Blue Microphone’s ‘Blue FiRe’ for iOS, Protect+ Voice Recorder for Android, and Audacity for Windows.

- **Digital voice recorder:** Voice recorders are specialised devices developed for use in interviews. Many provide multi-directional microphones and better audio pickup in comparison to smartphones, making them useful for focus groups. However, few offer security features to protect audio data captured in the field, and these are often expensive (costing £200+). Consult 1.4 for examples of encrypted voice recorders.

- **Laptop and local storage:** A laptop and USB microphone(s) may be used to record audio in a face-to-face environment using free software such as Audacity. Audio recordings held on portable storage should be protected using encryption software such as VeraCrypt, Windows BitLocker, or Apple FileVault.

- **Online conferencing software:** Skype, Skype for Business, GoToMeeting, Cisco WebEx, Facetime and other conferencing software are often used to perform interviews and focus groups where it is impractical to visit or gather participants. Researchers should ensure that their chosen communication method provides encryption and audio/text capture functionality.

- **Analogue voice recorders:** Analogue voice recorders that capture audio on cassette tape remain an option, but are strongly discouraged. If used, tapes should be stored and transported securely and transcribed at the earliest opportunity.
1.2. Configure audio settings

The configuration of the capture device has a significant impact upon the quality of the audio recording. When configuring the capture hardware/software, the following settings should be checked:

- **Sampling rate**: refers to the number of times that audio is recorded per second. This is measured in Hertz (cycles per second) or Kilohertz (thousand cycles per second). Voices can be recognised at a low sample rate such as 8kHz, however the use of a higher rate such as 44.1kHz or 48kHz is recommended on the basis it enables capture of a more complex sound wave, making it easier to understand an interviewee’s voice when recorded in a noisy environment.

- **Bit depth**: a measure of the number of bits of information in each sample – a higher number will produce a higher quality recording. Most software/hardware will set the audio bit depth without the need for user intervention, however in cases where it must be set, a value of 16 bit or higher should be used.

- **Number of channels**: The number of sound channels should ideally match the number of microphones used to record audio. If your audio recorder possesses one microphone only, a single-channel/mono recording is acceptable to record sound obtained through a single source. Many single microphone devices record in stereo by default, but this only replicates the same sound in the two channels. If you are recording a focus group discussion using multiple microphones, perhaps located in different parts of the table/room, the allocation of a larger number of channels would be useful. This will allow the researcher to isolate and listen to sound from a specific microphone.

- **File format**: The file format used to encode audio are often selected on the basis of the recording equipment in use. Dedicated voice recorders tend to support a small number of formats, typically MP3. If you use a smartphone or computer to record audio and have sufficient storage space, it’s advisable to use an uncompressed audio format such as WAVE (.wav), FLAC (.flac) or AIFF (.aif).

1.3. Ethics and informed consent

Research participants should be made aware of the data that will be captured, how it will be managed, and how it will be used during the study lifetime and following its completion.

Guidance on addressing data-related issues within the consent process can be found in ‘Informed consent for research’ Standard Operating Procedure (LSHTM SOP-005), located at https://lshtm.sharepoint.com/Research/Research-Governance/Pages/standard-operating-procedures-(sops).aspx

1.4. Data security in the field

The capture stage is a period of high risk, during which there is often only one copy of an audio recording. To comply with the LSHTM Information Security Policy1, the following steps to protect data in the field:

a. **Record only the audio that you need for research**

   You should only collect data that you require for research and for which you have been given consent. Recording devices should not be switched on before the interview/focus group begins.

---

1 https://www.lshtm.ac.uk/aboutus/organisation/information-management-and-security
and should be immediately turned off once it has concluded. It is good practice to check with participants that they understand the discussion will be recorded, gain agreement on when to begin recording and inform them when recording has been switched off. This is particularly important for telephone/online interviews, when the recording process is not always obvious.

Interviews that cover sensitive topics should be held in a private location where it cannot be overheard. A quiet location is encouraged, to ensure background noises are not recorded by the device.

You may wish to provide additional guidelines on information that should not be provided during the recording. For instance, "may I ask that you refer to family members by their relationship to you rather than by name?" The type of guidance is left to the interviewer's discretion; some interviewers may find it useful to guide the interview process to reduce the work needed to anonymise the transcript, whereas other researchers prefer to allow participants to express views without intervention and redact/anonymise information at the transcription stage.

b. **Use capture devices that offer built-in encryption:**
The use of an encrypted smart phone or digital voice recorder is required to protect your audio data in the event that the device is lost or stolen. Digital voice recorders that support 128/256-bit encryption include the Olympus DS-3500 Digital Voice Recorder, Olympus DS-7000 Digital Voice Recorder and Philips DPM8000. Many Apple/Android/Windows-based smartphones offer encryption functionality built-in, although this functionality may not enabled by default. Encryption tutorials for iOS and Android devices can be found in guidelines 4 and 5 of the Information Management and Security Policy\(^2\).

If a non-encrypted capture device is used, audio files must be transferred to a secure storage media, such as an encrypted laptop or the user’s LSHTM home drive immediately following the interview and the original copy securely deleted.

c. **Configure device security features:**
Many smartphones offer security features such as screen lock that requires a pattern/PIN/biometric feature to access the device, GPS tracking that can be used to locate the device, and remote wiping if the device is stolen, however these must be enabled by the user.

d. **Make a note of serial numbers and use a security marker to label devices**
Many devices provide serial numbers that allow them to be uniquely identified. For instance, a smartphone’s IMEI serial number can be found by typing *#06# into your handset. Security marker pens can also be used to add labels which can only be seen under UV light.

e. **Hide capture devices from view:**
Electronic devices should be hidden from view when being transported to reduce the risk they will become a target for theft.

f. **Store confidential resources in a secure location when not in use:**

\(^2\) [https://www.lshtm.ac.uk/aboutus/organisation/information-management-and-security](https://www.lshtm.ac.uk/aboutus/organisation/information-management-and-security)
Digital and physical resources containing confidential information should be stored in a secure area when not in use, such as a locked room or cabinet accessible only to yourself or a limited number of known individuals. Consult 2.1 for LSHTM digital storage options.

g. **Transfer audio data to a managed server at the earliest opportunity:**
Audio recordings and other data captured in the field should be uploaded to a managed server at the earliest opportunity. This will ensure they are held securely and backed-up on a regular basis. An encrypted laptop/portable device may be used in the absence of an internet connection, however this may itself become a target for theft.

If you possess a smartphone or laptop with internet capability, it is advisable to upload the files immediately following data collection. If an internet connection is unavailable, files should be transferred at the earliest opportunity, ideally by the end of the day.

h. **Remove audio files from the capture device as soon as possible:**
Audio recordings held on unencrypted devices represent a security risk. Once files have been uploaded to a managed server, the copies held on the capture device must be immediately deleted. Consult the ‘Data Destruction’ Standard Operating Procedure (LSHTM-SOP-043-01) at https://lshtm.sharepoint.com/Research/Research-Governance/Pages/standard-operating-procedures-(sops).aspx

For further advice on securing mobile devices before working in the field, consult the LSHTM ‘Lost Phone’ guidance at: https://lshtm.sharepoint.com/Services/IT-Services/info-security/Documents/lost_phone.pdf.

### 1.5. Reporting lost/stolen devices
Devices that contain personal data must be reported as lost/stolen at the earliest opportunity. The following steps should be taken as soon as the device is found to be missing:

1. Report the loss/theft to csirt@lshtm.ac.uk - the email address for potential information security incidents.
2. Report the loss/theft to the Police/British Transport Police to get a crime or loss reference number (for tracking/insurance).
3. If the device is synced with your LSHTM email or other accounts, change your password.

2. Prepare data for analysis

Several activities must be performed to prepare audio data for analysis. These include transcription, quality assurance, and anonymisation.

2.1. Data security

Data must be kept secure throughout the period that it is stored. Protection measures may include:

- Storing data in geographic regions that comply with the laws and regulations in the country/region in which it was collected.
- Use of managed storage systems that are regularly backed-up and protected by security measures that limit access to authorised users only, e.g. user accounts, encryption.
- Creation of an anonymised transcript for use during the analysis process.

Protection measures must be applied to *ALL* copies of the data, including primary and back-up storage. Guidance on LSHTM Data Storage Options can be found at:
Details of LSHTM requirements can be found in the Information Management and Security Policy, located at https://www.lshtm.ac.uk/aboutus/organisation/information-management-and-security.

2.2. Transcription

Audio recordings should be transcribed at the earliest opportunity. The transcript will often become the primary object for analysis, with the audio data being used for reference only from this point onwards. It is therefore important that a high-quality, accurate transcript is produced. Key decisions to be made at this stage include:

a. What audio should be transcribed?
b. What transcription conventions will be applied?
c. Who will perform the transcription?
d. How much time should be allocated and what resources are needed?
e. How will the accuracy of the transcript be established?

Further guidance is available through the following resources:

- UK Data Service: Transcription
  <https://www.ukdataservice.ac.uk/manage-data/format/transcription>

2.2.1. Select audio for transcription

The audio recording should be reviewed and a decision made on the amount and type of material that requires transcription. Is it necessary to transcribe the recording in its entirety or will a subset be sufficient?

2.2.2. Choose transcription convention

Second, consideration should be given to the transcription convention that will be applied. The Finnish Social Science Data Archive identifies several transcription levels:

1. *Summary transcription*: The key points and topics raised during the interview are noted and selected quotations recorded verbatim. The summary can be a useful resource for prioritising transcription.

---

activities, but are often insufficient for in-depth analysis or re-use due to the subjectivity of the information recorded.

2. **Be writ**Basic level transcription: An accurate transcript of the participants’ words and any significant expressions of emotion (laughter, sighs, etc.) are produced. However, statements not relevant to the discussion (e.g. words said when a participant answers a phone), non-lexical sounds (‘uh’, ‘ah’), and cut-off/repeat words may be left out.

3. **Exact transcription:** A verbatim, word-for-word transcription is produced, including fillers (‘you know’), repeats, cut-offs of words, non-lexical sounds, expressions of emotion (laughter, sighs, etc.) and emphasis or stress. Timed pauses (in seconds) and possible background noises and other disturbances are noted.

4. **Conversation analysis transcription:** The most detailed level of transcription. A full verbal transcript is produced using standard notation symbols, with careful reproduction of colloquial speech patterns. Transcription includes all words, timed pauses (in seconds), cut-offs of a word, intonation, volume, word stress, as well as non-lexical action (sneezes, breaths, sighs, facial expressions) etc.

There is often a trade-off in the amount of information needed to perform the research and the associated resources (time, cost) that can be allocated to perform the task. A basic level transcription (level 2) is often sufficient for many research purposes, with more detailed transcription performed to achieve specific research objectives.

2.2.3. **Estimate time required**

The time necessary to transcribe an audio recording should not be underestimated; it can take 4-7 hours to transcribe each hour of audio, depending upon the chosen transcription level, recording audibility, transcriber typing speed, language translation needs, and other factors. The Morgan Centre for Research in Everyday Lives offers an Excel-based tool to calculate time requirements: http://www.socialsciences.manchester.ac.uk/morgan-centre/research/resources/toolkits/toolkit-08/.

2.2.4. **Assign responsibility: In-house vs. outsource**

Transcription may be performed by a project member, research degree student, specialist transcription service, or other group/person hired for the task. Decisions to perform transcription in-house or outsource it are often influenced by factors such as available funds, expertise (language and domain knowledge), and availability at the relevant project stage.

If transcription is outsourced, the scope of the work and requirements should be clearly documented in a written contract and agreed prior to the commencement of work and transfer of funds. Key areas to cover include:

- No. of audio recordings to be processed, duration of each (in hours/minutes), source language, and other key characteristics
- Task to be performed, e.g. conversation analysis transcription, translation, etc.
- Confidentiality requirements, such as: audio must be held on an EU server, encrypted during storage and transfer, not passed to other parties without permission, etc.
The School does not maintain a list of recommended transcription services. Advice on suitable transcription services can be obtained by consulting with colleagues and reviewing the LSHTM Noticeboard at [https://lshtm.sharepoint.com/Services/Pages/noticeboard.aspx](https://lshtm.sharepoint.com/Services/Pages/noticeboard.aspx).

Contract and non-disclosure agreements for use of a transcription service must be written in conjunction with LSHTM's Legal Services team and approved by LSHTM's Research Operations office. The LSHTM Legal Services team can be contacted by emailing [legal@lshtm.ac.uk](mailto:legal@lshtm.ac.uk).

Data transfer to/from a transcription service must be performed in a secure manner. A tutorial outlining how 7-zip can be used to encrypt data prior to transmission can be found at [https://doi.org/10.17037/PUBS.03716462](https://doi.org/10.17037/PUBS.03716462). Advice on secure data transfer may be obtained from the Research Data Manager by emailing [researchdatamanagement@lshtm.ac.uk](mailto:researchdatamanagement@lshtm.ac.uk).

### 2.2.5. Locate software and hardware

Many generic tools exist that can be used for audio playback and transcription. However, it's worthwhile to explore specialised resources that may simplify the task. Examples of transcription software include:

- **ELAN Linguistic Annotator** - [https://tla.mpi.nl/tools/tla-tools/elan/](https://tla.mpi.nl/tools/tla-tools/elan/)
- **InqScribe** - [https://www.inqscribe.com/](https://www.inqscribe.com/)
- **Transcribe** - [https://transcribe.wreally.com/](https://transcribe.wreally.com/)

Voice recognition software that can be installed on a user's device, such as Dragon Naturally Speaking and MacSpeech Scribe, offer the potential to automate speech-to-text conversion, but often provide mixed results, misidentifying words and phrases and failing to identify multiple speakers unless they have been 'trained' to recognise a person's voice beforehand.

Researchers are discouraged from using cloud-based speech-to-text services. Although they provide greater accuracy in speech recognition in comparison to desktop-based tools, the terms and conditions for use of these services often state that they transfer and store data in geographic regions, which may breach terms set out in the participant informed consent agreement.

Hardware that may be useful to perform transcription include noise-cancelling head phones to block external sound and a USB Foot pedal for rewinding, fast forwarding, and pausing playback.

A technology watch on the use of transcription software will be maintained by the Research Data Manager and these recommendations will be reviewed and updated as necessary.

### 2.2.6. Quality Assurance

Transcripts should be reviewed and compared to the original recording to identify mistakes, such as missed and misheard words. This task should ideally be performed by someone other than the person who produced the transcript, such as a researcher who was involved in the interview/focus group, in order to provide an independent interpretation of the transcript.

To assist the review process, transcribers should be asked to label and timestamp sections of the recording that are inaudible or they cannot understand. For example: *[Inaudible between 00:45:20 – 00:52:53]* and *[Overlapping voices between 01:20:10 – 01:23:43]*.
2.3. De-identification and redaction

De-identification refers to the broad set of activities performed to prevent a research participant being identified when working with data. It is performed to enable a researcher to achieve one or more objectives. Common reason for performing de-identification include a need to:

- Enable data to be accessed by international collaborators without breach of data protection;
- Enable data to be stored on a specific server that can be accessed by the project team;
- Remove information that may influence the analysis;
- Enable data to be made available to researchers beyond the project team, in compliance with funding and journal requirements;

De-identification can be a time-consuming activity to perform, particularly for qualitative data. Therefore, it's important that you identify the reason it is necessary before allocating resources. Research studies that possess data that cannot be de-identified without context loss and a significant reduction in research value, for instance, may have justifiable reasons not to de-identify data. In these circumstances, it is essential that information security is maintained to protect data from unauthorised access and use at all times.

Consult the ‘Confidentiality and Anonymisation of Research Data’ Standard Operating Procedure (LSHTM-SOP-036) at https://lshtm.sharepoint.com/Research/Research-Governance/Pages/standard-operating-procedures-(sops).aspx for guidance. Information on secure locations to store confidential data can be found in section 2.1.

2.3.1. Audio recordings

Audio recordings contain a variety of identifiable information that may be used to identify the participant. In addition to the words that a person expresses vocally, vocal characteristics they provide - their speech pattern, accent, vocabulary - may be sufficient to determine their identity. Many techniques have been proposed to disguise personal & sensitive information, including the addition of a bleep censor sound effect and use of audio processing techniques such as dynamic pitch shifting and background noise addition. However, the time and effort needed to apply these protection measures can be significant. There is also potential that these protection techniques may be insufficient to disguise participants from machine learning techniques developed in the future. For this reason, the following approach is recommended:

a. If the participant has consented to being identified and has no objection to the audio recording being made available, e.g. the interview is performed to improve public knowledge, limited redaction - removing audio segments that cover sensitive or non-relevant topics - may be appropriate. However, no audio processing is necessary to disguise the participant's voice.

b. If the participant has not consented to being identified, the audio recording should be held in a secure location at all times and the transcript used as a basis for anonymisation and analysis.

2.3.2. Written transcripts

It's often considered easier to anonymise written transcripts in comparison to audio recordings, but it does not necessarily equate that it is a simple task. General rules to follow when de-identifying transcripts are:

1. Develop and apply a consistent approach to the handling of identifiable and confidential information:
   a. **Remove and replace:** Replace information with [brackets] or <tags> that denote edits (see below)
b. **Remove and store elsewhere:** assign a numeric identifier/code that can be linked to the confidential text segment and store the segment in a separate document held in a secure location. This will continue to be identifiable data until the identifiable content is deleted.

c. **Retain and record:** make a note of transcripts that contain identifiable and confidential information that must be retained and ensure these are held in a secure location;

2. Create an Anonymisation Log document in which you record details of all edits made to the transcript. The Anonymisation Log should indicate the type of information that has been removed, not the exact content.

3. If you must retain the raw, non-anonymised transcript for analysis, it should be stored in a secure location (such as the LSHTM Secure Server). This version *must not* be held on a public server accessible to the outside world.

4. Protection measures should be applied consistently to all data collected in a study, including qualitative and quantitative information collected across several waves of data collection during a longitudinal study.

Steps that may be taken to remove and replace identifiable and confidential information include:

- Replace identifying names and locations with pseudonyms (e.g. interviewee 1, organisation A) or labels that establish its role (e.g. father, shopping centre).
- Aggregate or reduce the accuracy of information, e.g. replace birth date with an age range (25-30), broaden geographic locations
- Disguise identifiable outliers by restricting upper or lower ranges, e.g. top-coding salaries.

Further advice on anonymising qualitative are available at the following resources:

- UK Data Service: Anonymisation of qualitative data
  [https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation/qualitative](https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation/qualitative)
- ICPSR: Confidentiality in qualitative data
  [https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter3qual.html](https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter3qual.html)

**2.4. File formats for analysis**

Review the documentation for your chosen software to determine the file formats it can import and export.

NVivo uses format encoder/decoders (known as ‘codecs’) installed on the user’s computer to import, playback and export audio content. This allows the use of a wide range of formats (such as MP3, MPEG4, WMA, WAV). However, the dependence on third party decoders rather than built-in functionality can cause problems when using NVivo on multiple devices; one researcher may be able to import MP3 content, whereas a collaborator who does not have an MP3 decoder installed will receive an error message when attempting to access it. If compatibility issues occur when using NVivo on your local machine, it is advisable to test it with the NVivo instance installed on LSHTM's Horizon virtual desktop.
2.5. Documentation

Documentation should be written throughout the research process that enables researchers – your future self and others - to understand the content and context in which the work was performed. Information that may support qualitative data include:

1. **Discussion guide**: The question list or topic list used by the interviewer to guide the discussion.

2. **Consent form and information sheet**: An unsigned copy of the informed consent form and information sheet provided to study participants.

3. **Data list**: A spreadsheet listing key characteristics on people interviewed during the one-to-one or focus groups discussions and the context in which the information was obtained. For example,
   - Interviewee ID
   - Interviewee age / age range
   - Interviewee gender
   - Interviewee occupation, organisation
   - Interview location
   - Interview date
   - Interview duration
   - Interview method
   - Language in which interview was performed
   - Language in which transcript was held
   - Key themes covered in the discussion
   - Interviewer name

Care should be taken to ensure documentation describing the research context does not provide information that may lead to the re-identification of participants.


3. Preparing for preservation

Research data, irrespective of whether it is qualitative or quantitative, are often subject to requirements on the time period it must be kept. These include:

- **LSHTM**: Funded and PhD data must be kept for at least 10 years following grant closure. MSc students are not covered by the LSHTM Retention and Disposal Schedule⁴, but should discuss data retention requirements with their supervisor.

- **Research funders**: Funders may specify a retention period of 10-25 years following grant closure, depending upon the funding call and type of research.

- **Journal publishers**: A growing number of journals expect data that underpins research findings to be available (in some form) for at least 10 years following the paper’s publication date.

---

• Country-specific legislation: Personal data contained in research data must not be kept for longer than is necessary to fulfil the intended purpose, in accordance with the UK Data Protection legislation.

These requirements may be interpreted in the context of qualitative research as follows:

a. Raw audio recording: Audio must be kept on the LSHTM Secure Server (or another secure environment) for the time period that they are needed. Once a transcript has been created and verified as accurate, an evaluation may be performed to determine if the audio recording continues to have research value and should be kept, or whether it would be better to delete it to protect research participants. If you are unsure on whether it should be retained, consult the Archives & Records Management Service (archives@lshtm.ac.uk) for advice.

b. Raw transcript: The un-anonymised transcript should be stored in a secure location for the time period that it is needed. If an anonymised derivative has been produced and the raw transcript is unlikely to be re-used, the un-anonymised version may be deleted on grant closure.

c. Anonymised transcript: The anonymised transcript can be kept permanently in many circumstances, unless there are third party requirements that require it to be deleted.

3.1. Preservation formats

Qualitative data should be held in one or more file formats likely to be accessible in the long-term. Researchers are encouraged to use open formats and standards supported by many software tools, where feasible. The use of ‘lossy’ compression formats, such as MP3, which reduce audio fidelity to save storage space, are discouraged, but accepted if it is the only file format supported by the digital audio recorders.

<table>
<thead>
<tr>
<th>Recommended preservation formats</th>
<th>Acceptable preservation formats (if originally held in this format)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUDIO DATA</td>
<td></td>
</tr>
<tr>
<td>Broadcast Wave Format (.bwf, .wav)</td>
<td>MPEG-1 Audio Layer 3 (.mp3)</td>
</tr>
<tr>
<td>Waveform Audio File Format (.wav)</td>
<td></td>
</tr>
<tr>
<td>Free Lossless Audio Codec (.flac)</td>
<td></td>
</tr>
<tr>
<td>TRANSCRIPTS</td>
<td></td>
</tr>
<tr>
<td>Plain text - Unicode, ASCII (.txt)</td>
<td>Hypertext Mark-up Language (.html)</td>
</tr>
<tr>
<td>Open Document Text (.odt)</td>
<td></td>
</tr>
<tr>
<td>Microsoft Office Open XML document (.docx)</td>
<td>Microsoft Word document (.doc)</td>
</tr>
<tr>
<td>Rich Text Format (.rtf)</td>
<td>NVivo (.nvp)</td>
</tr>
</tbody>
</table>

Table 1: Recommended formats for preservation
4. Prepare data for sharing

Qualitative research has traditionally been excluded from data sharing requirements. However, these expectations have changed in recent years. Research funders such as the Economic and Social Research Council (ESRC) now expect applicants to outline their approach to providing access to qualitative data within the research bid and enforce compliance as a condition of grant sign-off, while a growing number of journals require authors to state how data may be obtained in a data sharing statement.

Current data sharing expectations are influenced by the European Commission, which advocate a principle that research data should be "as open as possible, as closed as necessary". This encourages researchers to make research data available in some form for independent review: openly if ethical and legal consent have been provided, but recognises that some data cannot be made available to others due to the presence of personal, sensitive or confidential information, or may only be shared subject to additional security measures being applied.

4.1. Identify data to share

Researchers should consider the possibility that qualitative data will need to be made available to others during a study's planning phase, to ensure adequate time and resources can be allocated.

Sharing decisions should take into account the research purpose, ethical obligations, and legal framework in which the study is taking place. Key questions to consider include:

a. **What digital and physical resources underpin the research?**

   Consider the digital and physical resources that you will obtain use at each stage and how they contribute to the research. This may include: [1] a question list produced to guide the interview/focus group discussion; [2] participant consent forms and information sheets; [3] data in the form of audio recordings and written transcripts [4] consortium agreements and other legal documents; and [5] additional documentation that describe the data content and context in which it was obtained and analysed.

b. **What permissions must be obtained to permit other researchers to access and use these resources?**

   Research participants, research collaborators, and other relevant stakeholders must provide explicit consent for the resources to be shared before they can be made available. Researchers are expected to address sharing requirements in the participant informed consent form, consortium agreement and any other relevant legal agreement. Contact the Research Governance and Integrity Office for advice on informed consent and consult section 1.3 for further details.

   Projects that did not address data sharing as part of participant informed consent, but which are subsequently expected to make anonymised data available by a journal or funder must contact the LSHTM Research Ethics Committee for guidance.

c. **What resources can be made available? What resources must be withheld?**

   Health research involving human participants, particularly that performed in an international setting, are often subject to complex data protection, intellectual property and confidentiality regulations that dictate the data management and sharing strategy. Resources should only be

---

shared if explicit consent has been provided and appropriate measures have been applied to protect personal and other confidential data that must be withheld (see 2.3 for redaction guidance).

Qualitative research projects often share resources that enable others to understand how the investigation was performed. For instance, question lists, blank consent forms and information sheets, coding strategy, processing scripts produced to analyse data, and other administrative documentation.

The amount and type of research data that can be made available will vary between projects, influenced by factors such as the research topic, type of participant informed consent provided, ability to de-identify data, and so on; some projects may be able to share anonymised/redacted transcripts under controlled conditions, while others may only be able to provide selected quotes within the context of a research publication. Audio recordings of interviews and focus groups should not be made available, except in circumstances where participants have permitted identification and agreed to wider sharing (e.g. to improve public knowledge), as outlined in 2.3.1.

For advice on balancing the conflicting expectations that apply to research data, email the LSHTM Research Data Manager at researchdatamanagement@lshtm.ac.uk.

4.2. Select a digital repository to curate and preserve the data

Qualitative data may be deposited with one of several digital repositories, each of which are capable of handling the digital curation, preservation, and access process on the researchers’ behalf. These include:

- UK Data Service - https://www.ukdataservice.ac.uk/
- Qualitative Data Repository - https://qdr.syr.edu/
- LSHTM Data Compass - http://datacompass.lshtm.ac.uk/

Consult the Registry of Research Data Repositories (re3data) for other options at http://www.re3data.org/search?query=qualitative. For further advice on digital repositories contact researchdatamanagement@lshtm.ac.uk.

4.3. Oversight of the decision-making process for data access

A decision to be made when sharing restricted data is who will be responsible for handling the day-to-day tasks associated with reviewing access requests and taking decisions. Health researchers often build a rapport with research participants and wish to protect information they provide by taking custodial responsibility for evaluation of requests to access and use the data.

Research data may be kept for several years, which presents the possibility that researchers involved in the original research are no longer able to be involved in the decision-making process. Most digital repositories allow current data custodians to designate an individual with appropriate expertise to replace them.

4.4. Choosing an access method

Access methods for sharing qualitative data should be appropriate to the ethical and legal context in which the research was performed. The following approach represents two possible options:

1. Anonymised transcripts that can be shared without restriction may be made openly available under a Creative Commons Attribution (CC-BY) licence
2. Transcripts that cannot be fully anonymised or are covered by IPR that restrict how it may be used might be made available through an application process (i.e. on request), with usage restrictions specified in a Data Transfer Agreement (see 4.6).

4.5. **Determine permitted and non-permitted uses**
Qualitative data containing information obtained on the basis that it can be used only for specific purposes must be protected by a licence agreement that establishes conditions for allowed/non-allowed use. Contact the LSHTM Research Operations Office or Research Data Manager (researchdatamanagement@lshtm.ac.uk) for a copy of LSHTM’s sample Data Transfer Agreement.

4.6. **Adopt file formats suitable for use**
Qualitative data should be stored in open formats supported by common software tools in current use by end users. Consult UK Data Service advice at [https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats](https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats).

4.7. **Provide support documentation**
Documentation should be provided alongside qualitative data to enable researchers to understand the content and context in which the work was performed, without the need to contact the data creators. Key information to make available alongside a transcript include:

- **Discussion guide**: The question list or topic list used by the interviewer to guide the discussion.
- **Consent form and information sheet**: An unsigned copy of the informed consent form and information sheet provided to study participants.

Care should be taken to ensure documentation describing the research context does not provide information that may lead to the re-identification of participants.


**Definitions**

- **Encryption**: Encryption is a method of protecting digital information that works by scrambling the content of a file at the bit level. This is more secure than simple password protection, which does not change the arrangement of the file, making it possible to view in a hex editor. To unscramble the contained files, a user must provide the correct password. As a minimum, a 256 bit encryption algorithm should be used, such as AES 256.