



The work is licensed under the Creative Commons Attribution-Non-Commercial-Share Alike 2.0 UK: England and Wales Licence. To view a copy of this licence, visit creativecommons.org/licenses/by-nc-sa/2.0/uk/

Qualitative Data Collection Processing Procedures

Version 3.0

PUBLIC VERSION

07 October 2010

T +44 (0)1206 872001

E help@esds.ac.uk

www.data-archive.ac.uk



UK DATA ARCHIVE
UNIVERSITY OF ESSEX
WIVENHOE PARK
COLCHESTER
ESSEX, CO4 3SQ

WE ARE SUPPORTED BY THE UNIVERSITY OF ESSEX, THE ECONOMIC AND SOCIAL RESEARCH COUNCIL, AND THE JOINT INFORMATION SYSTEMS COMMITTEE

Contents

1.	Qualitative data	2
2.	Pre-processing checks on qualitative data collections.....	2
3.	Qualitative data formats	3
4.	Processing interview transcripts	3
4.1.	Making processing copies of original files.....	3
4.2.	The interview template	3
4.2.1.	Standard header information.....	4
4.2.2.	Transcript template font	4
4.2.3.	Interviewer/Respondent demarcation tags	4
4.2.4.	Formatting of transcripts and non-transcript items	6
4.3.	Processing transcripts	6
5.	Confidentiality and anonymisation of textual data	6
6.	User guides and additional metadata.....	8
7.	The Data List	8
7.1.	Confidentiality of the Data List	9
8.	Read and Note files	9
9.	Data file naming conventions for qualitative data.....	10
9.1.	Interview Data	10
9.2.	Focus Group Data	10
9.3.	Photographic and Video Data	10
10.	File naming conventions for ESDS Qualidata documentation	11
10.1.	National Social Policy and Social Change Archive (NSPSCA) documentation	11
11.	Digitisation of paper collections	11
11.1.	Digitisation and scanning	12

Scope

This document covers the processing of qualitative data collections archived at the UK Data Archive. It does not cover quantitative data processing, which is covered in the document *UKDA-DSS Quantitative Data Processing Procedures*. It refers to, but does not cover fully, documentation processing procedures, which are included in the document *UKDA-DSS Documentation Processing Procedures*.

It should be noted that some of the documents referenced within the text below are not publicly available, but external readers may of course contact the Archive in case of query.

1. Qualitative data

Qualitative data collections may contain a variety of materials, but currently the majority of deposits comprise a set of interview transcripts and accompanying documentation. Therefore, the instructions given in this document refer largely to interview transcripts, though the file naming conventions also cover other items such as audio (generally interview recordings) and image files. Techniques for working with audio, video and image files are still developing.

In the case of mixed-methods studies, the data collection may also include some quantitative data files. Instructions for processing quantitative data are covered in a separate document, *UKDA-DSS Quantitative Data Processing Procedures*.

2. Pre-processing checks on qualitative data collections

Once all administrative materials have been received, the Acquisitions section will pass the red folder to Qualidata for assessment. Qualidata staff (usually the Senior Qualitative Data and Support Services Officer) will check the data, documentation and administrative materials, and draw up a processing plan based on a template with recommendations for processing.

The data collection will then be passed to the Data Services section. Further preliminary checks may then be made by the Data Services Manager, and the processing plan annotated as necessary. A unique study number will be allocated, the materials placed into the standard archival directory structure, and the Calm processing database entry created. The data collection will then be passed to the Data Services team and full processing can then commence.

Occasionally, the Senior Qualitative Data and Support Services Officer will pass the data collection directly to a member of the Data Services team by prior arrangement.

Before processing commences, the processor should make a copy of the data collection in its archival structure on their allotted network processing area. The processing plan should then be read alongside the red folder contents.

3. Qualitative data formats

Most qualitative data collections currently comprise sets of interview transcripts, most commonly in Rich Text Format (RTF), Microsoft (MS) Word (.doc files), or plain text format (.txt). For further information on how to process sets of interview transcripts in these common formats, see section 2.2 of this document onwards. However, specialist qualitative software package files are sometimes included in the deposit.

Qualitative software packages such as CAQDAS, NUD*IST, ATLAS-ti and WinMax have export facilities that enables one to save a whole 'project' consisting of the raw data, coding tree, coded data and associated memos and notes. For archival purposes the raw data, the final coding tree and any useful memos should be exported prior to deposit. Depositors are requested to do this at the acquisition stage and checks should be made prior to processing to ensure it has been done. Coded data are not preserved, as they cannot as yet be exported in a common non-proprietary format. A set of procedures for processing ATLAS-ti format data collections is included in the document *Atlas Processing Guide*.

At present, coded data are not in demand, mainly because the coding process is subjective, often geared towards specific themes, and therefore may not be applicable to the secondary analyst's topic of investigation. For larger studies, there is a stronger case for retaining coded data, in order to aid searching within and navigation through voluminous bodies of text.

4. Processing interview transcripts

4.1. Making processing copies of original files

Before processing begins, good data management and archival practice dictates that a separate processing copy of all data and documentation files **must** be made, and that all processing work is done using these copies rather than the original files. The copies can be edited to fix errors and resolve confidentiality issues (see section 5 below), whereas the original files should be left in their original state, in case of future query. Original files are placed in the data collection structure under noissue/original/, with subdirectories as appropriate, and are not disseminated to users. Capitalisation and spaces should be removed from the original filenames before plattering (where a large number of files are affected, file renaming software is available for this purpose - please consult the Data Services Manager for details).

4.2. The interview template

A large proportion of qualitative digital data collections are deposited in MS Word format. However Rich Text Format (RTF) is the standard UKDA preservation format for this kind of text, and qualitative data are also typically distributed in RTF, for better cross-platform usability. To produce the standard Qualidata interview transcript format, an RTF interview transcript template should be constructed. This ensures that the correct font specification and speech demarcation is applied, and that standard header information is included.

4.2.1. Standard header information

Standard header information should be added to the template, which comprises the study number and title (may be copied from the unpublished catalogue record entry in the catalogue input programs), along with the depositor's name. The text file name (created at UKDA) should also be added, but this will of course vary according to the individual file. For file naming conventions, see section 9 below.

Example of standard header information:

Name of project: SN 6047 Masculinities in Transition: Identity, Home and Workplace, 2004-2006
 Depositor: Hockey, J.
 Filename: 6047int001

4.2.2. Transcript template font

The following fonts should be used throughout the interview transcript template: Verdana 11pt for the body and Verdana 10pt for headers and footers.

4.2.3. Interviewer/Respondent demarcation tags

Where the interview transcript follows the conversation recorded at interview, the text should be clearly separated into sections to show the interviewer's questions/statements and the respondent's responses, in the interview order.

With regard to naming (tagging) of the sections, the following procedure should be used. In the first instance, if the depositor (and his/her researchers) has been consistent and replaced personal names in the transcripts with pseudonyms or other acceptable substitutions or tags (e.g. 'Interviewer' and 'Respondent' or initials I or R or R1, R2, etc.), they may be left as deposited, with UKDA formatting applied.

However, if there is no consistency between transcripts, the tags should be changed so that they are consistent throughout the study. If no anonymisation has been done by the depositor, or the deposited version is unclear and inconsistent, change the tags to the initials I (for Interviewer) and R (for respondent). If there is more than one Interviewer and/or Respondent, for example where a multiple-respondent focus group has been conducted, numbered tags such as I1, I2, R1, R2, etc. may be used.

The tags should be in 11pt bold Verdana font. This will ensure that they are the same size as the standard transcript font, but emphasised by the bold setting. Standard formatting should be applied to the whole transcript (see section 4.2.4 below).

Some examples of the kind of tags that may be used are given below.

1. Where the depositor has used Interviewer/Respondent:

Interviewer: Text

Respondent: Text

2. Where the depositor has not used the Interviewer's name, but has used a pseudonym for the Respondent:

Interviewer: Text

Mary: Text

3. Where the depositor has used the Interviewer's first name and a pseudonym for the Respondent:

Simon: Text

Mary: Text

4. Where the depositor has used initials for both interviewer (SJ in this example) and respondent (MH in this example). (Check the study materials to ensure the use of the respondent's initials does not compromise their confidentiality.):

SJ: Text

MH: Text

5. Where the depositor has used initials for both Interviewer and Respondent (only one Respondent in this example):

I: Text

R: Text

6. Where the depositor has used initials for both Interviewer and Respondent (one Interviewer and two Respondents in this example):

I: Text

R1: Text

R2: Text

7. Where the depositor has used initials for both Interviewer and Respondent (two Interviewer and two Respondents in this example – increase numbers as appropriate for the study):

I1: Text

R1: Text

I2: Text

R2: Text

4.2.4. Formatting of transcripts and non-transcript items

The following formatting should be applied to interview transcripts (in MS Word, these settings may be found from the top menu under Format>Paragraph):

Alignment: Justified
Indentations: L = 0, R= 0
Special: Hanging by 3.3cm

The following formatting should be applied to interview notes or non-transcript items:

Alignment: Justified
Indentations: L = 0, R= 0
Special: (none)

4.3. Processing transcripts

Each deposited transcript may then be opened and copied into the interview transcript template, and the template saved as an RTF file with the appropriate UKDA naming convention (see section 8 below). Any header information from interview transcripts should be copied into the Data List (gender, age, location, date of interview, etc.).

Most of the actual processing of qualitative data involves familiarisation with the data that has been deposited. Reading the transcripts provides comprehension and understanding of the data and helps inform the drafting of the catalogue record.

The following checks should be made:

- the standard font specification (Verdana 11pt for the body and 10pt for headers and footers) is correct, and the hanging indent set is correct;
- double line spacing should be removed, and also extra lines or page breaks;
- special characters may need to be substituted or removed (e.g. " should be substituted for ");
- a spell-check should be run on the file to pick up obvious spelling and grammatical mistakes (i.e. your/you're; its/it's).
- Interviewer/Respondent tags (see section 4.2.3 above) should be inserted where necessary using 'Find and replace' and by reading through the transcript.

The transcript should then be read thoroughly to check for:

- items that could be added to enhance the data list (e.g. details of employment, education (suitably anonymised));
- logical consistency (e.g. 'find and replace' errors);
- correctness of formatting;
- confidentiality (see section 5 below).

During this thorough reading of each transcript, notes should be made on potential suitable index terms for the list of keywords.

5. Confidentiality and anonymisation of textual data

At the same reading of each transcript within the study, thorough confidentiality checks are made. Although UKDA users sign a legally binding access agreement to re-use data, and in that promise to respect guarantees of anonymity, consistent with the original investigator's undertaking, no information that clearly

breaches the confidentiality of the respondent or any other person or entity should be present in the dissemination version of the data collection.

Confidentiality is of paramount importance and depositors are thus requested to edit material prior to deposit. The task of the processing officer is to check this has been done well and consistently. For example checks should be made for errors where identifiers have been left intact or where so much material has been removed that sections of the data no longer have any meaning. In some cases it can be very difficult to disguise the identity of participants without introducing an unacceptable distortion into the data, and so full anonymisation may be impossible. Where problems are encountered, or the processor is unsure on how to proceed, advice should be sought from the Data Services Manager, the Senior Qualitative Data and Support Services Officer or the ESDS Social History Data Manager. Alternative solutions may be available, such as the restriction of user access to certain interviews within the data collection.

In some cases, exceptions to the normal confidentiality rules may be permitted. For example, permission may have been gained from respondents at the time of the original interview for re-use without anonymity, such as research with elites or life story material.

Guidance for depositors on pseudonyms and anonymisation techniques for qualitative data may be found at <http://www.data-archive.ac.uk/create-manage/consent-ethics/anonymisation?index=2>

The level of anonymisation to be adopted for any one data collection depends on the nature of the study and specifically on the consent agreements devised by the researchers. A key stage in acquisition and processing is the comprehension of the researchers' consent agreements. These are the foundation reuse lies upon. They also underpin the balancing of access restrictions versus anonymisation. As a result the specific procedure to adopt has to vary on a case-by-case basis. More detail on the approach to take is usually included in the individual data collection processing plan. If information has been deleted or altered for confidentiality reasons, a note should always be made in the Note file, and if likely to affect users, in the Read file too (see section 8 below).

In practice the main task undertaken to preserve confidentiality is the removal of major identifying details, i.e. real personal names, place and company names, street names etc., and replacement of them with pseudonyms where appropriate. It must be noted that very detailed information on employment/workplaces, educational institutions/qualifications, occupations of other family members and small geographical locations could all compromise confidentiality, even without revealing the respondent's name. On the other hand removing too much detail can lead to distortion of the data or a loss in accuracy.

Points to note are as follows:

- automated search and replace techniques may be used (such as MS Word's 'Search and Replace' function), but additional proofreading should always be carried out as automated processes are not foolproof;
- pseudonyms and pseudo place names should be the same as those used in any prior publication by the depositor/principal investigator;
- a file for internal UKDA use only, containing details of cross-referencing for pseudonyms to the original names (i.e. a 'key') should be compiled. This may be archived under 'noissue', and must not be disseminated to users.

However, as archival procedures dictate that the original version of each file deposited should be preserved intact (see section 4.1 above), the disclosive information should be retained in the original files, which remain under 'noissue' and are not disseminated to users.

Where any major problems are encountered, in particular citation of third parties, please consult the Data Services Manager or Senior Qualitative Data and Support Services Officer.

Note that care should also be taken to maintain confidentiality within the User Guide and Data List created for each data collection. Take note that this documentation is accessible to online users before registration and in that sense is public. In addition it is a resource that in some cases attracts more usage than the actual data; see sections 6 and 7 below.

6. User guides and additional metadata

User guides are compiled during processing for both qualitative and quantitative data collections – see the separate document *UKDA-DSS Documentation Processing Procedures* for full information on techniques used. For qualitative data collections, user guides generally comprise documentation or sometimes metadata relating to the data collection that have been supplied by the depositor, for example interview schedules, or End-of-Award report to the funding body. These files are checked for errors or breaches of confidentiality (such as email addresses or similar information) before they are then grouped together into a user guide (the exact nature of documentation/guides produced may vary with individual studies).

Note that the arrangements for header information and bookmark hierarchies are different for qualitative data collections to those for quantitative studies. Full instructions are given in the document *UKDA-DSS Documentation Processing Procedures*.

The reliance on using material supplied by the depositor means content can vary from collection to collection. Where qualitative data collections are processed to A* standard, and may be added to the Qualidata Online system, more metadata may be created, such as interview glossaries or the addition of commentaries on the research project. However this type of material is usually created either before or after conventional processing by the Qualidata team, though the processing officer may also be involved.

7. The Data List

A *Data List* is prepared for each qualitative data collection, to help users identify particular types of interviews or transcripts (such as women of a particular age in the sample, respondents from a certain location and so on). The Data List lists the key demographic characteristics of interviewees that define the sampled population, such as year of birth/age, gender, and perhaps geographical region. Depositors are routinely asked to collate and supply this information themselves prior to deposit of the data collection, which they may already have done in the course of conducting fieldwork or analysing data. The elements to be listed depend upon what has been recorded by the original researcher and your own reading which may pick out trends, so most listings will vary.

The Data List also serves other purposes. The list should indicate where data are missing, or if there is some variation; such as a mixture of interview transcripts and focus group transcripts, fully transcribed files or summary notes and so on. In addition, each file should carry a *unique identifier*. To ensure consistency during processing, the Data List should be started with the first interview transcript processed, and kept up to date as each transcript in the set is worked on. The Data List may be refined and edited throughout processing, depending on what information is available or relevant. Once all interview transcripts have been processed, the Data List should be finalised, with any redundant columns in the MS Excel template deleted. It is important to adhere to the template and not vary style. This provides one of the few opportunities to produce documentation that is consistent across a wide range of collections.

The use of MS Excel to create the Data List is helpful for larger data collections, as it enables the user to subset, search and filter if they wish, for example, to find those interviewees with certain demographic characteristics. However, an Adobe PDF version of the Data List is always created alongside the Excel version, to improve cross-platform usability. Therefore, the Excel Data List spreadsheet should be formatted to ensure that it is suitable for conversion to PDF - i.e. all text within cells is visible, no cells 'overhang' the page, and relevant columns are repeated across pages. Where the data collection contains many transcripts, this may require careful work. When formatting is complete, the Data List should then be converted to PDF, which can be done within Excel.

In some instances (e.g. lists with >100 cases) the spreadsheet may be too large to be fully readable in PDF format, and may be disseminated as Excel only. However, advice in these cases should be sought from the Data Services Manager and Senior Qualitative Data and Support Services Officer.

7.1. Confidentiality of the Data List

Pseudonyms may be included in the Data List as appropriate. Real names should not be included. See section 5 above for details of pseudonymisation, and the compilation of a 'key' document.

Information included in the Data List should be detailed enough to enable subsetting and filtering, but not detailed enough to enable respondent identification, either alone or in combination with other categories. The Data List is displayed on the UKDA website alongside the user guide, so confidentiality must be preserved. To use an extreme and hypothetical example of identification by category combination, an 'Orthopaedic surgeon' living in a small named village with a 'PhD in Spinal Surgical Techniques' may be quite unique and therefore potentially identifiable; a 'medical practitioner' in the same village with a 'postgraduate degree', is less identifiable, but information on their occupation and the level of their qualifications is still largely intact. Classification schemes (such as SOC2000 for occupations) may be useful for this purpose.

Sometimes, two versions of the Data List may be produced. For example, where some but not all of the interviews in the data collection are restricted to permission-only use, a separate Data List should be created for the restricted set, which will not be displayed on the web, but will be supplied with the restricted interviews to those users who gain permission to use them.

Secondly, where the sample is large, the Data List may need to contain more detail in order to enable useful filtering and subsetting. In this case, a less detailed version of the Data List should be created for the web, and the more detailed version confined to the download package created for dissemination to users. It is currently under consideration whether this dual Data List system should be put in place for all data collections.

If the processor has any queries concerns regarding confidentiality within the Data List, the Data Services Manager, the Senior Qualitative Data and Support Services Officer or the ESDS Social History Data Manager will be happy to advise.

8. Read and Note files

As for quantitative data collections, two HTML format metadata files, called Read and Note files, are compiled for each qualitative data collection during processing. They are held in the Calm processing database, though their structure differs slightly between qualitative and quantitative studies. Both files contain information about processing history - checks carried out, problems discovered, confidentiality edits made, etc., but are created for different purposes, which must be borne in mind when deciding what information to include:

- the Read file is for external display on the UKDA website, and is distributed to the user with the data collection download package;
- the Note file is for internal use only.

For qualitative data collections, information should be added regarding work carried out on the data collection, e.g. clear notes on anonymisation techniques used and the replacement of identifiers. As the Read file is visible to users, it must not contain any confidential information that may have been included in the Note file, such as keys to replaced names. There is also a particular need for mixed methods data collections to include information on whether and how elements inter-relate and may be linked between quantitative and qualitative files (e.g. how a case in an SPSS questionnaire file may link to the same respondent's RTF interview transcript). Full procedures for creation of Read and Note files are given in the document *Creating Read and Note files in Calm V1_1* (soon to become *UKDA-DSS Creating Administrative Metadata*).

9. Data file naming conventions for qualitative data

9.1. Interview Data

In most cases file naming is based on a descriptor of the **event** (interview, focus group etc) and numerical identifier prefixed by the appropriate study number. It will **not** change according to format.

Example: labelling Interview Number One in SN 2000:

- RTF text transcript: label as 2000int001 (an RTF file, will be delivered as 2000int01.rtf)
- An MP3 audio file: label as 2000int001 (an mp3 file, will be delivered as 2000int01.mp3)
- A WAV audio file: label as 2000int001 (a wav file, will be delivered as 2000int01.wav)

Where a single interview is represented by one transcript but a number of audio recordings – e.g. because the interview was conducted over a number of visits but transcribed as a single file – then additional letters a,b,c, etc are added for the audio.

Example: SN 4890 ('Madness in its Place') includes an interview (no.27) with a former patient, comprising a single RTF transcript file but also a number of audio files:

- RTF text transcript: 4890int027
- Three audio recording files: 4890int027a
4890int027b
4890int027c

A similar approach can be used for any event within a research project defined as occurring a number of times, but which remains related, e.g. a longitudinal project which uses successive interviews with a respondent but still defines them as a single event.

Note: For 'legacy' reprocessing of older qualitative data collections in the collection, the renaming of files to current conventions may cause significant work. File renaming software is available for this – please consult the Data Services Manager for details.

9.2. Focus Group Data

A focus group interview is a different kind of event to the individual interview, so the transcript filename will reflect that change:

- SN 4890 Focus Group Transcript One: 4890fg001

9.3. Photographic and Video Data

Still images or video footage of a general nature (and unrelated to a particular interview or event) need to reflect this fact with a different label:

- SN 4890 Still Image Number One: 4890pic001
- SN 4890 Video Clip Number One: 4890vid001

However, if the image was related to a specific event (such as picture of a focus group or a particular person interviewed) it would be named to reflect that event:

Example: the fifth image in a group of ten images specifically related to Interview Number Eight would be named:

- SN 4890 Still Image Number Eight (Five of Ten): 4890int008e.tiff (where 'e', the fifth letter of the alphabet, denotes the fifth image)

File names will vary in response to research events (i.e. interview transcript, field notes, a character sketch, classroom observation, etc.) and appropriate labels are decided by discussion between the processor and

the Senior Qualitative Data and Support Services Officer. The key question is the nature of the research event to be labelled. It is also important that once a new labelling term is introduced this is re-used when another collection has that kind of data event. It is the responsibility of the processor to ensure this.

A further illustrative example, this time from an imaginary study, is:

SN 2234 - Teaching Poetry

- Poem: label as 2234poem001
- Field notes: label as 2234notes001
- Notes/ transcription of classroom observations: label as 2234obs001
- Examples of writing: label as 2234writ001

Note: where only one recording of an individual research event (e.g. an interview transcript) exists, the format suffix (e.g. .rtf) is not usually included in the Data List. However, where additional formats of the same event exist (e.g. an audio recording of the same interview), and are to be made available and listed in the Data List, the format suffix should be included for both files to avoid confusion.

10. File naming conventions for ESDS Qualidata documentation

The naming conventions for ESDS Qualidata documentation differ slightly from those for UKDA quantitative study documentation (see separate document UKDA-DSS Documentation Processing Procedures for details of those). Qualidata documentation is named as follows:

- User Guide (Adobe PDF format): label as XXXXuguide (where XXXX is the study number, e.g. 2234uguide)*
- Data List (Excel and PDF): label as XXXXulist

Multiple documentation volumes are not usually a feature of qualitative data collections (though there are exceptions). However, PDF files should be kept below 10mb in size, so where a great deal of documentation does exist for a qualitative data collection, it is better practice to organise files by the type of material included, e.g. '2234methodology', '2234reports' etc., as this is clearer for the user, and also in keeping with the subject-based documentation work carried out at the UKDA by the Survey Question Bank (SQB) project. Where extensive documentation exists, advice on processing and naming should be sought from the Data Services Manager or the Senior Qualitative Data and Support Services Officer.

*Documentation for older qualitative data collections may have a 'q' prefix (e.g. 'q2234uguide'), but this is no longer necessary.

10.1. National Social Policy and Social Change Archive (NSPSCA) documentation

Data Lists for studies from NSPSCA which may have been added to the UKDA catalogue do not follow the usual Excel format, and may instead be labelled:

- Paper archival list*: label as 2234palist

If processors are required to create a virtual catalogue record and Data List for an NSPSCA data collection, advice should be sought from the Data Services Manager or the Senior Qualitative Data and Support Services Officer.

11. Digitisation of paper collections

ESDS Qualidata have previously undertaken in-house preservation of paper materials, but this is now rarely done. However, paper collections that are selected to be worthy of digitisation (in whole or in part) are stored and then prepared for digitisation by Qualidata staff. This forms an important part of the added value aspect of material processed to A* standard. Most of the time-consuming tasks involved with this actually take place before normal processing begins. The intention is to create a collection of digital files that can then be

passed onto Data Services for processing and archiving. Further metadata may also be produced at this stage (see section 6 above).

The nature and form of paper materials means a number of issues for digitisation have to be considered:

- suitability of the material for digitisation (content, paper quality; type or handwritten);
- the proportion of the collection to be digitised;
- how the collection should be prepared with respect to physical, organisational or intellectual considerations;
- to what extent text should be made machine-readable (i.e. TIFF image or some level of Optical Character Recognition (OCR));
- what level of improvement of the resulting images should be performed.

The most critical consideration is whether the files resulting from the digitisation process should be simply stored as images (i.e. a TIFF image of the paper) or whether they should be converted to fully searchable text (i.e. OCR is performed). Due to the complex nature of qualitative data collections, which can include printed paper questionnaires and schedules with typed and hand-written comments, some materials may not be suited to OCR.

11.1. Digitisation and scanning

For materials that contain poor typeface, handwriting, tables or drawings, the paper is scanned and saved as an image, in Tagged Image File Format (TIFF) file format. For each 'document' or transcript, all the constituent TIFFs are grouped together using Adobe PDF, to preserve the 'look' of the original paper. The PDF file is then bookmarked and may be annotated. Adobe PDF security settings may also be applied to the files where necessary. Full details of image improvements, PDF conversion and bookmarking should be set out in detail in the Note file.

Edits may still be made to hard copy material for reasons of confidentiality (see section above). This should be performed prior to scanning. For example, any page of an interview transcript that contains sensitive material should have the relevant sections obscured with black permanent marker. Original copies must NEVER be marked in this manner; the material to be edited should be photocopied and the marker pen used on that. The edited photocopy may then be scanned.

Once the full set of digitised documents has been produced, the data collection should be processed in the normal way. All transcripts must be proof-read and checked for confidentiality, and a Data List compiled.