QUALITATIVE DATA COLLECTION INGEST PROCESSING PROCEDURES

PUBLIC
16 APRIL 2014
Version 07.00

T +44 (0)1206 872001
E sharonb@essex.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

Scope

This document covers the ingest processing of qualitative data collections archived at the UK Data Archive. It does not cover quantitative ingest processing, which is covered in the document *Quantitative Data Ingest Processing Procedures.* However, where mixed-method data collections contain both qualitative and quantitative data elements, the procedures in this document and *Quantitative Data Ingest Processing Procedures* must be followed as appropriate. Finally, this document refers to, but does not cover fully, Documentation Ingest Processing Procedures, which are included in the document *Documentation Ingest 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), video and image files. In the case of mixed-methods studies, the data collection may also include some quantitative data files. As noted above, procedures for processing quantitative data are covered in *Quantitative Data Ingest Processing Procedures*. It may be possible to link quantitative and qualitative data elements to respondents within the same study, for example a survey file may exist alongside a set of interview transcripts. This should be explicitly mentioned in the information provided to users with the collection.

2. Pre-processing checks on qualitative data collections

2.1. Processing Plan

Once all administrative materials have been received, the Collections Development team will pass the materials to the Ingest Services team for assessment. The data, documentation and administrative materials will be checked and a unique study number will be allocated, the materials placed into the standard archival directory structure, and the Calm processing database entry created. In addition a processing plan is drawn up (based on a template) with recommendations for processing. The processing plan is produced as an aid to processing staff. This is based on knowledge of the collection and deposit negotiations and an initial review of the material after deposit. In addition, Collections Development and Producer Support staff may have been involved in assisting the depositor with preparing the data for deposit, suggesting anonymisation strategies and advising on access restrictions. The Ingest Services team will consult with the other teams as needed.

Production of the processing plan will involve;

- Final assessment of completeness and quality of the deposit
- Assessment of confidentiality issues and wishes of depositor (e.g. access conditions)
- Advice on the ingest processing tasks to be done (layout edits, data listing, user guides)
- Notes on unusual details - such as early issue of study numbers, handling of audio or images

The ingest processing demands of qualitative collections vary considerably depending on the individual collection. Ingest staff are therefore encouraged to see the plan as flexible and open to discussion. Its purpose is to help orientate the processing officer to a collection and provide a clear statement on its size, content, access conditions and individual requirements. Once the processing plan has been created, the collection is ready for ingest processing. Before processing commences, the processor must make a copy of the data collection on their allotted network area and work on that, so there is always a ‘master’ copy available during ingest processing, in case of problems.

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 below. However, specialist qualitative software package files are also sometimes included in the deposit.

Qualitative software packages such as Nvivo, NUD*IST, ATLAS-ti and MaxQDA have export facilities that enable one to save a whole 'project' consisting of the raw data, coding tree, coded data and associated memos and notes. The coding process is often subjective and geared towards specific themes, and therefore may not be useful for the secondary analyst's topic of investigation. However, for larger studies, coded data may be helpful to aid searching and navigation through voluminous bodies of text. For archival purposes the raw data, the final coding tree and any useful memos should be exported as separate files prior to deposit. The coded elements of the collection can be provided to users as long as raw data are also made available. Depositors are requested to export all coded data to export formats prior to deposit and checks should be made prior to processing to ensure this has been done. It may not be possible to reliably preserve coded and annotated data for the long-term at the Archive, as they cannot all be exported in a common non-proprietary format (though some packages do use XML to export coded data, which is suitable for archiving). A set of procedures for processing ATLAS-ti format data collections is included in the document *Atlas Processing Guide* (not currently a controlled document).

4. Ingest processing of interview transcripts

The processing of qualitative material is an exercise in comprehension. The need for careful checking of content means it is not sufficient to read a sample. All data and documentation is read to protect the interests of participants, data creators, data users and the UK Data Archive. Whilst this need is unavoidable, wherever possible ways of standardising stages in ingest processing work are sought. This allows for efficiency of time and consistency in how a diverse range of collections can be presented.

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 (unless the depositor requires their destruction). 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).

4.2. Formatting the interview transcript

If the depositor has provided consistently clear and well-formatted interview transcripts throughout the set, these may be retained with the minimum of editing being undertaken. Depositors should be encouraged to set transcripts to the standard Archive format prior to deposit. A transcript example is available for their use. However, a pragmatic approach is required if this is likely to prove difficult for the depositor, for example where the project has long since ended and project staff have moved to other posts.

A large proportion of qualitative digital data collections are deposited in MS Word format. However Rich Text Format (RTF) is the standard Archive preservation format for this kind of text. Qualitative data are typically distributed in RTF, for better readability than plain text and cross-platform usability. Therefore, while the original Word files are of course retained for archival storage, an RTF copy of all MS Word files must be created for ingest processing, dissemination and archival purposes, using the appropriate Archive naming convention (see section 9 below).

Editing and formatting of text are normally kept to a minimum during ingest processing, as this can be very time-consuming and resource-intensive. Where the interview transcripts are in reasonable format as received from the depositor, editing will be limited as long as demarcation between the interviewer's questions and the interviewee's responses is clear. However, where the data collection is subject to enhanced processing, for example if it is part of a special project such as the *Pioneers* data collection, an RTF interview transcript template using the standard Archive interview transcript format may be constructed. This will ensure that consistent font specification and speech demarcation is applied, and that standard header information is included.

4.2.1. Standard header information

The Archive's standard header information 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. for brevity, this is enough for most standard collections, as long as the interview id number is displayed at the head of the interview. The text file name (created at the Archive) may also be added for enhanced collections. For file naming conventions, see section 9 below. Where the depositor has already added the bulk of the information in a consistent format and font throughout the set of transcripts, the Archive-specific information only need be added.

Example of enhanced header information

Name of Project: SN 9999 Mothers' Relationships with their Teenage Daughters, 2010
Depositor: Bloggs, J.

Interview ID: M001

Filename: 9999int001

4.2.2. Transcript template font

If the depositor has provided well-formatted and clear fonts consistently throughout the interview transcripts set, these may be retained. The Archive's standard font for interview transcripts is: Verdana 11pt for the body text and Verdana 9pt for headers and footers. Depositors should be encouraged to format transcripts in this font prior to deposit.

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.

If the data collection transcripts deposited already include a consistent method of identifying interviewers and respondents that has been applied consistently across the set of interviews, e.g. first names (where permitted), pseudonyms or initials, these may be used. If identification has not been applied consistently, a logical method of tagging **that suits the particular collection** may be chosen and used for all transcripts within the study. Some examples of the kind of tags that may be used are given below.

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 below).

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

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

Simon: Text

Mary: 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 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

4. Where the depositor has used Interviewer/Respondent:

Interviewer: Text

Respondent: 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

If the depositor has used clear formatting consistently throughout the interview transcripts set, this may be retained. Where this has not been applied, the Archive's standard format for interview transcripts is as follows:
Alignment: Justified
Indentations: L = 0, R= 0
Special: Hanging by 1.27cm

Depositors should be encouraged to format transcripts in this manner prior to deposit.

Note that where tags/pseudonym names are long, the tab size may need to be adjusted. It should be kept to the minimum value that provides consistent indent for blocks of speech, to retain maximum readability for the data user.

The Archive's standard format for interview notes or non-transcript items is as follows:

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

If the depositor has used clear formatting consistently throughout the notes/non-transcript item set, this may be retained. Depositors should be encouraged to format non-transcript items in this manner prior to deposit.

4.3. Reading and ingest processing transcripts

Most of the actual processing of qualitative data involves comprehension and 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 Data List and Discover catalogue record. The data creator will have used content for their own research needs but the role of the processor is to also consider other, perhaps unused, potential of the data that can be highlighted in the documentation.

The transcript should be read thoroughly to check for:

- (suitably-anonymised) items for the data list (demographic information such as gender, age, location, date of interview, etc.).
- 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 overall study themes covered that may be useful for keyword indexing.

The following edits should be made:

- if the standard Archive font has not been used, where the font used by the depositor has not been applied clearly and consistently, the standard font specification (Verdana 11pt for the body and 9pt for headers and footers) should be applied
- if the standard Archive hanging indent set has not been used, where the indents used by the depositor have not been applied clearly and consistently this should be corrected;
- speaker identification tags (see advice given in section 4.4 above) should be inserted where necessary.
- double line spacing may need to be removed, and also extra lines or page breaks;
- special characters may need to be substituted or removed.

A spell-check may need to be run on the file to pick up obvious spelling and grammatical mistakes (i.e. your/you're; its/it's).However, where transcript errors and spelling mistakes are widespread, it may be too

resource-intensive to correct them. The errors should be left in situ and notes made in the Note and Read file where appropriate. (Please be sensitive in the wording used in the externally-available Read file.) Remove time coding, marginalia, annotations, highlighting or any other markings left by the depositor as part of their analysis. Note: **Do not** correct errors produced by a dialect or idiomatic speech.

5. Confidentiality and anonymisation of textual data

At the same reading of each transcript, thorough confidentiality checks are made. Although Archive 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. Initial suggestions regarding this will be made in the processing plan.

Confidentiality is of paramount importance and depositors are expected to anonymise all data and documentation prior to deposit. The processing work undertaken at the Archive should be confined to checking, not basic anonymisation. The task of the processing officer is to check that anonymisation has been done well and consistently. For example checks should be made for errors where identifiers that should have been anonymised prior to deposit have been left intact or where so much material has been removed during pre-deposit confidentiality editing 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 Curation 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 (retrieved November 17, 2014).

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 that data sharing and re-use 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 and should be guided by the depositor (e.g. to enable the same pseudonyms to be used in the data as in any publications based on data). 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 this can introduce additional errors. Further proofreading should always be carried out to check the results after using it;
- 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 Archive 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. If necessary, access control measures such as password protection may be used on the files as an extra precaution.

Where any major problems are encountered, please consult the Data Curation Manager.

Note that care should also be taken to maintain confidentiality within the User Guide and Data List created for each data collection. It should be remembered that these documents are accessible to online users before registration and in that sense are public. If necessary, separate document versions for the website (less detailed) and for use with the data by registered users (more detailed) may be compiled - see sections **Error! Reference source not found.** and 7 below.

6. Documentation metadata

Documentation metadata, sometimes including user guides, are compiled during processing for both qualitative and quantitative data collections – see the separate document *Documentation Ingest Processing Procedures* for full information on techniques used. For qualitative data collections, user guides have generally comprised information relating to the data collection that have been supplied by the depositor, for example interview schedules, or an End-of-Award report to the funding body. However, it is desirable nowadays to keep different kinds of documents (such as methodology, research reports, etc.) as separate volumes, to enable more sophisticated analysis and search retrieval by services such as Question Bank.

Documentation files must be checked for errors or breaches of confidentiality (such as email addresses, over-detailed funding information or similar information) and possibly edited before conversion to Adobe PDF format and possible combination with other documents.

Note that the arrangements for header information and bookmark hierarchies may differ for qualitative data collections to those for quantitative studies. Full instructions are given in the *Documentation Ingest 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 made available online, more metadata may be created, such as interview glossaries or the addition of commentaries on the research project. These extra resources are not often created, but the Producer Support team should be able to provide further details where enhanced processing is to take place.

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. On past occasions, summaries of the topics covered in the interviews may have been included, but this is no longer done– demographic detail is sufficient.

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 columns 'overhang' the page, and relevant column headings 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 in 'landscape' format, which can be done within Excel (checking in 'Print Preview' before conversion gives a good idea of how the resulting PDF file will look).

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 Curation Manager.

7.1. Confidentiality of the Data List

Information included in the Data List should be detailed enough to enable subsetting and filtering by respondent characteristics, but not detailed enough to enable respondent identification, either alone or in combination with other categories. The Data List is displayed on the Archive website and Discover record 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 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. Mapping information to standard 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 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 authenticated users.

- Where the depositor has provided a detailed Excel data list-type document. All depositors are encouraged to do this before deposit – see http://www.data-archive.ac.uk/create-manage/document/data-level?index=2 . This may be used as the basis for a Data List and the detail retained in a version for the data collection package that registered users will download, and **a separate version should be created for the web catalogue with detail removed.**

- 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.

- 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.

If the processor has queries or concerns regarding confidentiality within the Data List, the Data Curation Manager should be consulted.

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 via Discover and is distributed to the user with the data collection download package. It should contain advice or useful tips on using the collection;
- the Note file is for internal use only and will generally be much more detailed.

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.

8.1. Providing Read file information for mixed-methods data collections

For mixed methods data collections, information may need to be added to the Read file on whether and how elements inter-relate and linkage 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).

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 1 in SN 2000:

RTF text transcript: label as 2000int001 (an RTF file, will be delivered as 2000int01.rtf)MP3 audio file: label as 2000int001 (an mp3 file, will be delivered as 2000int01.mp3)

- FLAC audio file: label as 2000int001 (a .flac file, will be delivered as 2000int01.flac)

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 (representing the interview) but also a number of audio files (since it was recorded on several of tapes)

- 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' processing of older qualitative data collections in the collection, the renaming of files to

current conventions may cause significant work. File renaming software can be used for this, e.g. – the Bulk Renamer software currently installed on Ingest Services computers.

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 1: 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 1: 4890pic001
- SN 4890 Video Clip Number 1: 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 8 would be named:

- SN 4890 Still Image Number 8 (5 of 10): 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.). 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. New labels can be added to this document to ensure continuity.

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 qualitative documentation

10.1. User Guides

The naming conventions for qualitative documentation may differ slightly from those for the Archive's

quantitative study documentation (see separate document *Documentation Ingest Processing Procedures* for details of those), though documents are usually presented separately for later search retrieval purposes.

However, where an old-style combined user guide is to be used, it should be named as follows:

- User Guide (Adobe PDF format): label as XXXXuguide (where XXXX is the study number, e.g. 2234uguide)[1]

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 Question Bank's subject-based documentation work

10.2. Data Lists

The Data List should be named as follows:

- Data List (Excel and PDF): label as XXXXulist (where XXXX is the study number, e.g. 2234ulist)

10.3. Virtual catalogue records and National Social Policy and Social Change Archive (NSPSCA) documentation

Data Lists for studies from the NSPSCA (a paper based archive maintained by the University of Essex Albert Sloman Library) and other archives, which may have been added to the Discover catalogue in the past, do not follow the usual Excel format, and may instead be labelled:

- Paper archival list*: label as 2234palist

Virtual catalogue records are no longer routinely created, but if processors are required to create one, advice should be sought from the Data Curation Manager.

11. Digitisation of paper collections

In-house preservation of paper materials has been undertaken by the Archive in past times, and has also recently been undertaken under the Digital Futures project for inclusion in the Discover Qualibank system (see http://discover.ukdataservice.ac.uk/QualiBank). Paper collections that have been selected to be worthy of digitisation in whole or in part (most commonly classic sociology collections) were stored and then prepared for digitisation. This forms an important part of the added value aspect of material processed to the enhanced A* standard. The resulting files can then be passed onto the Ingest Services team 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.

[1] Documentation for older qualitative data collections may have a 'q' prefix (e.g. 'q2234uguide'), but this is no longer used for new collections.

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 and the option of rekeying may be considered.

When the decision is taken to digitise a paper collection, a set of guidelines is created in order to maintain consistency through the whole process; as digitisation can be time consuming, and more than one staff member may undertake the work. In addition, some collections may also include non-paper materials such as audiotape recordings of interviews, which require digitisation by specialised methods and equipment.

11.1. Digitisation and scanning

For materials that contain poor typeface, handwriting, tables, photographs 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. Whilst the Ingest processing officer may not do the work they should be familiar enough with what was done in order to add appropriate comments to the Read and Note files.

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.