



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/](http://creativecommons.org/licenses/by-nc-sa/2.0/uk/)

# Data Processing: Quick Reference

Version 2.0

---

PUBLIC VERSION

12 October 2010

---

T +44 (0)1206 872001

E [help@esds.ac.uk](mailto:help@esds.ac.uk)

[www.data-archive.ac.uk](http://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.	<b>'Calm' processing database</b> .....	2
2.	<b>Pre-processing review</b> .....	2
3.	<b>Allocation of unique study number</b> .....	3
4.	<b>Data processing</b> .....	3
4.1.	Quantitative studies.....	3
4.1.1.	Quantitative studies in SPSS format.....	3
4.1.2.	Quantitative studies in formats other than SPSS.....	4
4.2.	Qualitative studies .....	4
4.3.	Mixed methodology data collections .....	4
5.	<b>Documentation processing</b> .....	4
6.	<b>Read and Note files</b> .....	5
7.	<b>Naming of files and checking of study directory structure</b> .....	5
8.	<b>Red folder directory (rf)</b> .....	5
9.	<b>Creating the label file (.lbl)</b> .....	5
10.	<b>Preparing the study/data collection for download</b> .....	6
11.	<b>Archiving the study/data collection on the preservation system</b> .....	6
12.	<b>Checking the archived study</b> .....	7
13.	<b>Cataloguing and indexing</b> .....	7
14.	<b>Creating variable list(s) and frequency displays</b> .....	7
15.	<b>Red folder scanning and storage</b> .....	7
16.	<b>Checklist of main ingest processing activities</b> .....	8

## Scope

### What is in this guide?

This document is a brief but comprehensive overview of the main stages of ingest processing undertaken on studies/data collections acquired by the UK Data Archive/Economic and Social Data Service (ESDS). It is intended as an introductory guide for new Data and Support Services staff. New staff will normally undertake ingest processing work under the guidance of a fully trained Data and Support Services Officer. Certain stages as outlined below will remain the responsibility of the supervising Officer until training is complete.

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.

### What is not covered by this guide?

This guide does not contain fully detailed data and documentation ingest processing procedures. These are covered in the documents *UKDA-DSS-Quantitative Data Processing Procedures*, *UKDA-DSS-Qualitative Data Processing Procedures*, *UKDA-DSS-Documentation Processing Procedures* and *UKDA-DSS-Data Processing Standards*. This document **does not** cover processing procedures for Secure Data Service (SDS) studies/data collections.

## 1. 'Calm' processing database

'Calm' is an internal Archive database, used for recording study processing progress. Separate instructions on how to use CALM are available elsewhere. New staff should familiarise themselves with CALM and ensure that all study processing records are kept up to date.

## 2. Pre-processing review

When a new study/data collection is deposited with the UK Data Archive, it is first received by the Acquisitions section. Once initial administration is complete, the study is passed to the Data Services (DS) section for ingest processing. All incoming studies are reviewed by the Data Services Manager or another senior DS member prior to full ingest processing.

The pre-processing review includes (but is not limited to) the following checks:

- completeness of study/data collection (i.e. whether all materials have been received)
- data and documentation confidentiality
- whether documentation coverage is sufficient
- selection of processing standard (aka 'level of processing')

Following the review, a set of processing notes is written, which should be read thoroughly before processing commences. If the data collection is qualitative, the review will be in the form of a processing plan document, which will have been written by the Senior Data Services Officer (Qualidata) and augmented by the Data Services Manager.

In most cases, the unique study number will already have been allocated, and the study structure (see below) and Calm database entry created at the pre-processing review stage. If not, see below for appropriate procedures.

### 3. Allocation of unique study number

Prior to the allocation of the unique study number (often referred to as 'booking in'), studies are referred to by their acquisition number, allocated when the Acquisitions team add the study to the acquisitions database, 'Mirage' (also accessible from the Calm interface). Allocation of the study number is done by a member of the DS section trained in cataloguing, and is carried out using the Archive's catalogue record input program.

The unique study number appears in the web catalogue record for the study, whereas the acquisition number is used for internal administrative purposes only. After the study number has been allocated, the study is generally referred to by the study number rather than its acquisition number. 'Study number' is normally abbreviated to 'SN' elsewhere at the Archive, and in the remainder of this document.

## 4. Data processing

### 4.1. Quantitative studies

In practice, the vast majority of quantitative microdata files are deposited in SPSS format, and it is also by far the most popular dissemination format. Processing a quantitative study therefore typically entails:

- converting the data into SPSS .sav format where appropriate, if that is not the deposit format;
- performing integrity and validation checks on the data according to its processing standard (A\*, A, B, or C);
- creating dissemination and preservation formats (usually SPSS, Stata and tab-delimited text).

The required integrity and validation checks undertaken are described in the *UKDA-DSS-Data Processing Procedures* and *UKDA-DSS-Data Processing Standards* documents.

Once the validation and integrity checks are complete (and any addition/edits of variable and value labels undertaken) and the SPSS .sav files are ready for secondary use, they must be placed in the SN/spss/spss12 directory (where SN = Study Number) within the study structure. It is likely that this will have been done at the pre-processing review stage, but there may be exceptions. Once the study structure is correct, the SPSS processing script can then be run, as described below.

#### 4.1.1. Quantitative studies in SPSS format

The UK Data Archive currently uses a processing script written in-house to create alternative dissemination/preservation formats and internal metadata files. Full instructions on how to install the script and run it may be found in the documents *UKDA-DSS-Processing Script Procedures* and *UKDA-DSS-Data Processing Procedures*.

The script runs in SPSS and automates creation of the following data and metadata files:

- Stata format data files;
- tab-delimited text files (.tab);
- a warning log file of data loss or truncation from SPSS to Stata (due to the differential data handling limits of the two packages);
- creation of data dictionary files (generated from the SPSS data dictionary);
- creation of preservation data (fixed-width ASCII) and metadata files.

#### 4.1.2. Quantitative studies in formats other than SPSS

If the study is one of the few that is not deposited in SPSS format and/or is not processed in SPSS format (e.g. MS Access), the relevant procedures may be found in the *UKDA-DSS-Data Processing Procedures* document.

## 4.2. Qualitative studies

Processing qualitative data collections typically entails performing integrity and validation checks on the data according to its processing standard (A\* or A), developing a data listing and creating dissemination formats. Most qualitative data collections currently acquired take the form of individual interview, or focus group, transcripts. Details of relevant procedures may be found in the document *UKDA-DSS-Qualitative Processing Procedures*.

Most qualitative material is received in Word or Rich Text Format (RTF) format, and is made available as RTF. In the (increasingly rare) cases where qualitative material is provided as Tagged Information Format File (TIFF) files (usually old collections where paper documents have been scanned), the TIFF files are grouped together as Adobe Portable Document Format (PDF) files. For material in MS Access format, consult the relevant information in the *UKDA-DSS-Data Processing Procedures* document.

## 4.3. Mixed methodology data collections

Mixed methodology data collections include both quantitative and qualitative elements (e.g. a set of Word interview transcripts and an SPSS survey file). For these collections, a combination of qualitative and quantitative processing should be used, depending on the nature of the materials contained in the data collection.

## 5. Documentation processing

Alongside data files, the user will need documentation to help them understand the research project and analyse the resulting data files. Accompanying documentation takes many forms, but for survey data usually comprises a questionnaire, technical report, methodological information, and details of data variables. For qualitative data collections, it may comprise an interview schedule and methodological information.

Most documentation is supplied in electronic format, usually in Word, RTF or Excel. Some may be scanned into TIFF files from hard (paper) copy, though this is rare nowadays. Most documentation processing comprises conversion of these files to Adobe Acrobat PDF format, and adding bookmarks and headers to aid navigation. Several files may be combined into one or more user guide volume(s) or they may be left as individual files, depending on the size of the files and the nature of the study. Multi-worksheet Excel files may be left as they are if unsuitable for PDF conversion.

**Note:** TIFF files from scanned hard copy documentation are retained for archival purposes, so do not delete them after conversion to PDF.

Full documentation processing procedures are contained in the document *UKDA-DSS-Documentation Processing Procedures*, which also includes useful tips for those not familiar with Adobe Acrobat software. The Archive has naming conventions for documentation files. However, certain depositors like to retain the file names as supplied.

## 6. Read and Note files

Two metadata files, named 'Read' and 'Note' files, are compiled during study processing. They are held in the Calm processing database. Both files contain information about processing history - checks carried out, problems discovered, etc., but are created for different purposes - this must be borne in mind when deciding what information to include in each:

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

## 7. Naming of files and checking of study directory structure

The Archive has strict conventions on study structure, directory names and file extensions. These must be obeyed. The study structure is usually set at the time of the pre-processing review.

The universal rule on file naming is that spaces in file names should always be replaced, usually by an underscore. Characters such as ampersands '&' and brackets '()' should also be removed.

The Archive keeps electronic copies of all original files deposited with the study. The parent directory for storing these is <SN>/noissue/original. These file names are not usually altered, except for the removal of spaces and other forbidden characters.

## 8. Red folder directory (rf)

Each study will have a 'red folder' associated with it (Special Licence studies have a 'yellow' folder). The physical red folder contains all paper administrative documents concerning the study. An electronic red folder, or 'rf' directory is also created for each study, containing email correspondence undertaken with the depositor at the negotiation/acquisitions stage, and deposit/data review forms. The processing notes sheet accompanying each study should contain notes on the location of the electronic red folder directory.

## 9. Creating the label file (.lbl)

Once all data and documentation files and formats are complete, a text file is created that contains the name of each file for distribution to users. This file, which has the extension '.lbl', is used by the online documentation table in the Archive/ESDS catalogue, where available documentation files and their contents are listed. It is also used as the basis for an information file included in the download package (see section below).

The filenames and tab characters are created an internal program, which runs in UNIX. An example of a typical .lbl file is given below.

**Example of .lbl file:**

6052datadocs.pdf	Data Documents
6052interviewingdocs.pdf	Interviewing Documents
6052userguide.pdf	User Guide
wh_07_adults_archive.dta	Welsh Health Survey 2007 Adult Data
wh_07_adults_archive.sav	Welsh Health Survey 2007 Adult Data
wh_07_adults_archive.tab	Welsh Health Survey 2007 Adult Data
wh_07_adults_archive_UKDA_Data_Dictionary.rtf	UKDA Data Dictionary
wh_07_child_archive.dta	Welsh Health Survey 2007 Child Data
wh_07_child_archive.sav	Welsh Health Survey 2007 Child Data
wh_07_child_archive.tab	Welsh Health Survey 2007 Child Data
wh_07_child_archive_UKDA_Data_Dictionary.rtf	UKDA Data Dictionary
read6052.htm	UKDA Information for Study 6052
UKDA_Study_6052_Information.htm	Study information and citation

The structure of the file is as follows:

file name 1<tab>brief description of file contents of file 1

file name 2<tab>brief description of file contents of file 2

Running the .lbl file program will list the file names. Some automatic labels, such as the Read file and Information .htm file labels are added automatically, but the other labels will need to be added.

The .lbl file can then be opened in a text editor (UltraEdit or PFE packages are fine; Excel and Word are also useful, as long as the file is saved in plain text format). Type in the labels after each tab and then resave the file with the same name (<SN>.lbl).

An adequate description should be given for each file. There is no hard limit on the number of characters per label, though it should be kept at less than 60 characters unless there are good reasons why a more lengthy label is necessary. The Archive's standards and conventions on file labelling are available in a separate document.

## 10. Preparing the study/data collection for download

The Archive runs a 'download' system for the majority of its collection, where registered users can log in to their account, select the study they need, and download a zip file containing the study/data collection in the format of their choice (usually SPSS, Stata or tab-delimited for quantitative studies, and RTF for qualitative data collections, as per the formats created during ingest processing). Therefore, once ingest processing is complete, dissemination copies of the study are prepared for download system (these are separate to the copy held on the archival preservation system).

The download zip packages are prepared by running two scripts, one in SPSS and one in UNIX. Full instructions on how to create the download packages are given in a separate document. Note that the download script cannot be run until the catalogue record has been completed and 'published' (see below).

**Note:** Although in a small minority, some studies are either not in standard formats, or have restrictive access conditions. Their download packages are created manually, and may have an extra level of security put in place.

## 11. Archiving the study/data collection on the preservation system

When ingest processing is complete, the Digital Preservation and Systems team will copy it from the work area onto the Archive's preservation system. This is known as 'plattering'.

Before requesting that a study be 'plattered', a final check should be made that all file and directory names are correct, no temporary files have been accidentally left behind by any software applications or procedures,

and the study is in the approved study structure. If all is correct, a plattering request may be made through the ESDS IT HelpDesk. Once the plattering request has been entered, an automated email notification with the database job number will be sent to the member of staff who requested it.

## 12. Checking the archived study

The study **must** be checked once it has been plattered, to ensure that all is correct. Plattering problems do occasionally happen. An email notification that the HelpDesk job has been closed will be sent when the study has been plattered. It can then be viewed on the read-only 'front end' of the preservation system. The files will be deleted from the staff member's processing area once they have been plattered, so a copy should be kept in another location until plattering is complete and correct. Once plattering has been checked, this backup copy must be deleted, for reasons of data security (see document *UKDA Data Security Procedures*).

## 13. Cataloguing and indexing

Cataloguing and indexing of studies is a complex procedure, and as such it is usually the last part of training undertaken for DS processing staff and is best left until they become fully familiar with the intricacies of ingest processing.

For those staff who have not yet received specific catalogue training, once the study is plattered (or at the agreed stage), notification should be given to the designated member of the DS team and the red folder passed back to them so they can complete the catalogue record. For those DS staff who have received catalogue training, the catalogue record and keyword index should be completed according to the procedures and rules in the *Cataloguing Guidelines and Procedures* document.

Once the study has been plattered successfully, and the catalogue record and keyword index completed, the Data Services Manager or another member of the DS team will 'release' its catalogue record. The study will appear in the web catalogue and will become available for users to order.

Catalogue records are subject to pre- and post-release quality control checks: the procedures for this are outlined in the document *UKDA-DSS-Catalogue Quality Control Procedures*.

## 14. Creating variable list(s) and frequency displays

Variable lists and frequency displays are displayed on the web for each suitable, quantitative, SPSS format study; this functionality allows users to search variables via the Archive/ESDS catalogue. These displays are created by means of an in-house program that takes the information from the ddi\_xml file created by the processing script (see section above). The program will be installed once the member of staff concerned has received catalogue training.

## 15. Red folder scanning and storage

This section covers only the physical materials associated with red folders (or yellow folders for Special Licence studies). The electronic materials associated with red folders should be archived according to the procedures outlined in elsewhere.

After the study is released, the physical red folder will be passed to the member of the Acquisitions team responsible for scanning, and notification will be made separately that the electronic red folder is ready for archiving. The Acquisitions staff will then generate an email to notify depositors that the study has been released. They will also scan any administrative documents (correspondence, licence forms and any miscellaneous notes) that are not available in electronic format and archive the resulting image files. If any materials in the red folder are simply paper copies of files available electronically (e.g. deposit forms, data submission forms or documentation), they should be clearly marked 'Do not scan'. The Acquisitions scanning staff are also responsible for storing the physical red folders in the Safe Store.

## 16. Checklist of main ingest processing activities

New staff may find it useful to refer to this list for each study/data collection processed. A similar checklist is available on request for qualitative processing.

1. All data files have been processed and converted into suitable dissemination formats
2. All documentation has been created, with bookmarks and headers as appropriate
3. Read and Note files have been completed in CALM and saved as .htm format files
4. All files created during processing are appropriately named and located in the correct directory
5. Directory structure is complete and correct, including rf directory if appropriate
6. All original files received from the depositor have been copied to the relevant SN/noissue/original/subdirectories, with no changes to file names other than the replacement of spaces with an underscore and removal of &, (), etc.
7. Study has been prepared for download, and test zip packages checked
8. Study has been sent for plattering
9. Plattering has been completed and checked
10. Cataloguing has been completed
11. Keyword index has been completed
12. Catalogue record has been released
13. Where appropriate, variable lists have been uploaded to the catalogue record.
14. The CALM database has been updated
15. Red folder has been passed to scanning staff.