

## External Document

# Data Processing: Quick Reference

### UK Data Archive (UKDA)

Creator:	Data Services, UK Data Archive
Maintained by:	Data Services, UK Data Archive
Contributor:	
Version:	01.00
Controlled document No:	80
Last Amended:	2 March 2009
Review due date:	2 March 2010

Data Services  
UK Data Archive  
University of Essex  
Wivenhoe Park  
Colchester  
Essex  
CO4 3SQ

email: [help@esds.ac.uk](mailto:help@esds.ac.uk)  
Tel: +44 (0) 1206 872001  
Fax: +44 (0) 1206 872003  
[data-archive.ac.uk](http://data-archive.ac.uk)

## Table of Contents

<b>Document Control</b> .....	<b>2</b>
<b>Dataset ingest processing</b> .....	<b>3</b>
<b>1. The 'CALM' processing database</b> .....	<b>3</b>
<b>2. Pre-processing review</b> .....	<b>3</b>
<b>3. Allocation of unique study number</b> .....	<b>4</b>
<b>4. Data processing</b> .....	<b>4</b>
4.1. Quantitative studies .....	4
4.1.1. Quantitative studies in SPSS format .....	4
4.1.2. Quantitative studies in formats other than SPSS .....	5
4.2. Qualitative studies .....	5
4.3. Mixed methodology datasets.....	5
<b>5. Documentation processing</b> .....	<b>5</b>
<b>6. Read and Note files</b> .....	<b>6</b>
<b>7. Naming of files and checking of dataset directory structure</b> .....	<b>6</b>
<b>8. Red folder directory (rf)</b> .....	<b>6</b>
<b>9. Creating the label file (.lbl)</b> .....	<b>6</b>
<b>10. Preparing the study for download</b> .....	<b>7</b>
<b>11. Archiving the study on the preservation system</b> .....	<b>8</b>
<b>12. Checking the archived dataset</b> .....	<b>8</b>
<b>13. Cataloguing and indexing</b> .....	<b>8</b>
<b>14. Creating variable list(s) and frequency displays</b> .....	<b>8</b>
<b>15. Red folder scanning and storage</b> .....	<b>8</b>
<b>16. Checklist of main ingest processing activities</b> .....	<b>9</b>

## Document Control

Version	Notes	Last Amended
01.00	First external version	2009-03-02

### Replaces or supersedes:

*Data Processing Overview*

### Review terms:

Annual

### Is related to:

*UKDA Data Security Procedures*

*UKDA-DSS-Data Processing Standards*

*UKDA-DSS-Data Processing Procedures*

*UKDA-DSS-Documentation Processing Procedures*

*UKDA-DSS-DataProcessing-CALM Recording of Dataset Processing*

*UKDA-Study Structure (in development)*

*UKDA-DSS-Processing Script Procedures (in development)*

*UKDA-DSS-Processing-File Labelling Standards*

*UKDA-DSS-Processing-Download Package Creation*

*Cataloguing Procedures and Guidelines*

*UKDA-DSS-Creating Administrative Metadata*

*UKDA-DSS-Qualitative Processing Procedures (in development)*

**Note:** many of the above documents are currently under review or partly compiled from older documents. Please consult the Data Services Manager for current status.

## Scope

### What is in this guide?

This document is a brief but comprehensive overview of the main stages of UKDA dataset ingest processing procedures. It is intended as an introductory guide for new Data and Support Services staff. For external readers, it is for information only, and readers should note that some of the documents referenced may not be available for external use.

### 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-Data Processing Procedures*, *UKDA-DSS-Documentation Processing Procedures* and *UKDA-DSS-Data Processing Standards*.

## Dataset ingest processing

New staff will normally undertake ingest processing work under the guidance of a fully trained Data and Support Services Officer. Certain stages outlined below will remain the responsibility of the supervising Officer until training is complete.

### 1. The 'CALM' processing database

'CALM' is an internal UKDA database, used for recording dataset ingest processing progress.

Separate instructions on how to use CALM may be found in the document *UKDA-DSS-DataProcessing-CALM Recording of Dataset Processing*. New staff should familiarise themselves with CALM and ensure that all dataset processing records are kept up to date.

### 2. Pre-processing review

When a new dataset is deposited with the UKDA, it is first received by the Acquisitions section. Once initial administration is complete, the dataset (in its 'red folder') is passed to the Data Services section for ingest processing. All incoming datasets are reviewed by the Data Services Manager prior to full ingest processing.

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

- dataset completeness (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 study 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 have been allocated, and the study structure (see sections 7 and 8 below) and CALM database entry created at the pre-processing review stage. If not, see section 3 below and the documents *UKDA-DSS-Data Processing-CALM Recording of Dataset Processing* and *UKDA-Study Structure* for appropriate procedures.

### 3. Allocation of unique study number

Prior to the allocation of the unique study number (aka 'booking in'), studies are referred to by their acquisition number, allocated when the Acquisitions team add the study to their database, 'Mirage' (also accessible from the CALM interface). Allocation of the study number is done by a member of the processing team trained in cataloguing, and is carried out using the UKDA 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 dataset is generally referred to by the study number rather than its acquisition number. 'Study number' is normally abbreviated to 'SN' elsewhere at the UKDA, 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 appropriate place within the study structure. It is likely that this will have been done at the pre-processing review stage, but if not a UNIX script can be run to create the structure. Once the study structure is correct, the SPSS processing script can then be run (please consult the Data Services Manager for the current location of the SPSS and UNIX scripts).

#### 4.1.1. Quantitative studies in SPSS format

The UKDA 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* (in development) 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 (including ddi\_xml).

#### 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. Microsoft (MS) Access), the relevant procedures may be found in the *UKDA-DSS-Data Processing Procedures* document.

### 4.2. Qualitative studies

Processing qualitative data 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 ingested at the UKDA takes the form of individual interview, or focus group, transcripts. Details of relevant procedures may be found in the document *UKDA-DSS-Qualitative Processing Procedures* (in development).

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 studies 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 datasets

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

## 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, 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 PDF, 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 UKDA has naming conventions for documentation files.

However, certain depositors prefer that the original file names are retained - see the document *UKDA-DSS-Documentation Processing Procedures* for details.

## 6. Read and Note files

Two metadata files, called Read and Note files, are compiled during dataset ingest 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:

- the Read file is for external display on the UKDA website, and is distributed to the user with the dataset download package;
- the Note file is for internal use only. This must be borne in mind when deciding what information to include in each.

Procedures for creation of Read and Note files are given in the document *UKDA-DSS-Creating Administrative Metadata*.

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

The UKDA has strict conventions on dataset structure, directory names and file extensions. These must be obeyed. The dataset 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 UKDA keeps electronic copies of all original files deposited with the study. The original file names are not usually altered, except for the removal of spaces and other forbidden characters.

## 8. Red folder directory (rf)

Depending on the materials deposited, the dataset may also contain a directory named 'rf'. This is normally created at the pre-processing review stage (see section 2 above), and is used to store electronic versions of the deposit forms (XML or RTF formats, used to create the catalogue record - see section 11 below) and the data submission form (used by the UKDA Acquisitions Review Committee (ARC) to assess the dataset for addition to the collection). The rf directory should be plattered with the rest of the study, and is for internal use only. The filenames will not be picked up by the labels program, and should not be included in the .lbl or the download package information list (see sections 9 and 10 below).

The naming conventions for red folder files (if not added at pre-processing review stage) are as follows:

- XXXX\_depo.rtf (where XXXX = study number) - RTF deposit form
- XXXX\_depo.xml - XML deposit form
- XXXX\_formats.xml - XML formats form
- XXXX\_subm.rtf - RTF data submission form

MS Word format deposit or submission forms should be converted to RTF format before archiving.

## 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 UKDA 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 10 below).

The filenames and tab characters are created by the labels program that 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

The makelbl program will list the file names, and add some automatic labels, such as the Read file and Information .htm file labels, but the other labels will need to be added.

The .lbl file can then be opened in a text editor (UltraEdit or PFE 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 UKDA standards and conventions on file labelling are given in the document *UKDA-DSS-Processing-File Labelling Standards*.

## 10. Preparing the study for download

The UKDA runs a 'download' system for the majority of its collection, where registered users can log in to their account, select the dataset they need, and download a zip file containing the dataset in the format of their choice (usually SPSS, STATA or tab-delimited for quantitative studies, and RTF for qualitative, as per the formats created during ingest processing). Therefore, once ingest processing is complete, dissemination copies of the dataset are prepared for the 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 the document *UKDA-DSS-Processing-Download Package Creation*. Note that UNIX script cannot be run until the catalogue record has been completed and 'published' (see section 11 below).

**Note:** Although in a small minority, some datasets 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. Full details are given in the document *UKDA-DSS-Processing-Download Package Creation* (in development).

## 11. Archiving the study on the preservation system

When ingest processing is complete, the Digital Preservation and Systems team will copy it from the work area onto the UKDA 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 UKDA 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 dataset

The dataset **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 checked to ensure plattering has been successful. 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 the plattered version 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 datasets 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 UKDA ingest processing.

For those staff who have not yet received specific catalogue training, once the dataset 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 Procedures and Guidelines* 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 UKDA catalogue. These displays are created by means of a program that takes the information from the ddi\_xml file created by the processing script (see section 4.1.1 above). The program will be installed once the member of staff concerned has received catalogue training.

## 15. Red folder scanning and storage

After the study is released, the red folder will be passed to the member of the Acquisitions team responsible for scanning. They will then generate a letter to notify depositors that the study is released. They will also scan all loose documentation (correspondence, licence forms and any miscellaneous notes) 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 dataset 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 complete and correct, including rf directory if appropriate
6. All original files received from the depositor have been copied to the relevant 'noissue' 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. The CALM database has been updated
14. Red folder has been passed to scanning staff