DATA INGEST PROCESSING QUICK REFERENCE

PUBLIC
02 NOVEMBER 2015
Version 09.00

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

This work is licensed under the Creative Commons Attribution-ShareAlike 4.0 International Licence. To view a copy of this licence, visit http://creativecommons.org/licenses/by-sa/4.0/

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

Contents

Scope

What is in this guide?

This document is a brief but comprehensive overview of the main stages of data ingest processing undertaken on studies/data collections acquired by the UK Data Archive. It is intended as an introductory guide for new Ingest Services staff. New staff will normally undertake ingest processing work under the guidance of a fully trained Data Curation 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 *Quantitative Data Ingest Processing Procedures, Qualitative Data Ingest Processing Procedures, Documentation Ingest Processing Procedures* and *Data Ingest Processing Standards*. This document **does not** cover processing procedures for Secure Access studies/data collections, which are are documented in *Secure Access Ingest Procedures*.

1. CALM processing database

'CALM' is an internal UK Data Archive database, used for recording study processing progress. Separate instructions on how to use CALM are available. 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 Collections

Development section. Once initial administration is complete, the study is passed to the Ingest Services section for ingest processing. All incoming studies are reviewed by the Data Curation Manager or another senior member of Ingest staff prior to full 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') (This may also be subject to discussions undertaken at the negotiation/acquisition stage.)

Following the review, a processing plan is written, which should be read thoroughly before ingest processing commences. In most cases, the unique study number will already have been allocated at this stage, and the study structure (see below) and CALM database entry created. If not, see below for appropriate procedures.

3. Allocation of unique study number

Prior to the allocation of the unique study number, studies are referred to by their acquisition number, allocated when the Collections Development team create an entry in the acquisitions database, 'Mirage' (also accessible from the CALM interface). Allocation of the study number 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 large 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 (but not always) entails:

- converting the data into SPSS .sav format where appropriate, if that is not the deposit format;
- performing integrity and validation checks on each data file according to its processing standard (A*, A, B, or C). The required integrity and validation checks undertaken are described in the *Quantitative Data Ingest Processing Procedures* and *Data Ingest Processing Standards* documents.
- creating dissemination and archival storage packages (usually SPSS, Stata and tab-delimited text; others by prior arrangement);
- creating a suitable catalogue metadata record for the study.

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/spss19 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 and Python processing scripts 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 obtained from the Data Curation Manager.

The scripts currently run in Python within SPSS and automates creation of the following data and metadata

files:

- Stata format data files;
- tab-delimited text files (.tab);
- creation of data dictionary files;
- creation of preservation data (fixed-width ASCII) and associated preservation metadata files.

4.1.2. Quantitative studies in formats other than SPSS

If the study is not deposited and/or not processed in SPSS format (e.g. MS Access or other database/data formats), the relevant procedures may be found in the *Quantitative Data Ingest Processing Procedures* document.

4.2. Qualitative studies

Processing qualitative data collections typically entails performing integrity and validation checks on the data, 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 *Qualitative Data Ingest Processing Procedures.*

Most qualitative material is received in Word or Rich Text Format (RTF) format, and is made available as RTF.

For material in MS Access format, consult the relevant information in the *Quantitative Data Ingest Processing Procedures* document.

In the (now very rare) cases where qualitative material is provided as Tagged Information Format File (TIFF) image files (usually old collections where paper documents have been scanned), the TIFF files are grouped together as Adobe Portable Document Format (PDF) files.

4.3. Mixed methodology data collections

Mixed methodology data collections include both quantitative and qualitative elements (e.g. a set of Word interview transcripts plus a quantitative 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 derived/weighting 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. Most documentation processing comprises conversion of these files to Adobe Acrobat PDF format, and adding bookmarks and headers to aid navigation. Multi-worksheet Excel files may be left as they are if unsuitable for PDF conversion.

Combining documentation

Following technical developments undertaken for the Question Bank project during 2010/11, documentation is usually left in its separate component volumes for easy and precise user searching purposes. Older studies in the Archive collection may often contain combined documentation and combination of documents may still be done depending on the nature of the study or depositor requirements (for example the Health

Survey for England).

Paper documentation
Some documentation may be scanned into TIFF files from hard (paper) copy, though this is very rare nowadays, as most documentation is deposited in digital format.TIFF files are then converted to PDF, but the original TIFFs should be retained for archival purposes (archived under SN/noissue/mrdoc/tiff/, so do not delete them after conversion to PDF.

Full documentation processing procedures are contained in the document *Documentation Ingest Processing Procedures,* which also includes useful tips for those not familiar with Adobe Acrobat software. While the Archive has some naming conventions for documentation files, they are flexible and certain depositors may prefer to retain the file names as supplied (usually discussed on a case-by-case basis).

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, which must be borne in mind when deciding what information to include in each:

- the Read file is for external website display alongside the catalogue record, and is distributed to the user with the study/data collection download package;
- the Note file is for internal Archive 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)

An electronic directory (referred to as ‘rf’ or ‘red folder’) is created for each study. This contains email correspondence undertaken with the depositor at the negotiation/acquisitions stage, deposit/data review forms, and licence forms. Hard-copy licence forms and correspondence are scanned by the Collections Development team before the study is passed to the Ingest team for processing (see the *Pre-Ingest Procedures*. Sometimes, issues encountered during ingest processing necessitates correspondence between the Ingest team and the depositor. This correspondence, including the post-release email (see section 14), should be converted to PDF format and archived with the study.

Why ‘red folder’?
When hard copies of correspondence were the norm, physical cardboard folders were used for each study. From January 2013, these were phased out in favour of electronic versions. The physical folders were coloured ‘red’ for standard access/End User Licence access studies (hence the term ‘red folder’), ‘yellow’ for Special Licence studies, which are subject to more restrictive access conditions, and ‘green’ for Secure Data Service studies. Note that the ‘red folder’ or rf directory is still used as a file structure convention irrespective of access level. Also, physical yellow folders were marked ‘Protect’ according to the Archive’s *Information Classification Policy* meaning that its contents are subject to additional protection.

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 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 by a program named 'makelbl', currently run in the ZipDiss program. 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   UK Data Archive 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    UK Data Archive Data Dictionary
read6052.htm                                    UK Data Archive 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 labels, such as those for the Read file and Information file .htm files are added automatically, but the other labels will need to be added.

The .lbl file can then be opened in a text editor (NotePad ++ or PFE32 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 one tab space 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.

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 and are known by their OAIS name, 'Dissemination Information Packages', or DIPs).

The download zip packages are prepared by running two scripts, currently one in SPSS and one via the ZipDiss program. Full instructions on how to create the download packages are given in a separate document. **Note that the download script that creates the final DIPs, that runs on the Archive's preservation system after the study is plattered, cannot currently run successfully 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 include an extra level of security.

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

When ingest processing is complete, the archival information package (AIP) created for the study will be transferred to the Archive's preservation system for archival storage. 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 JIRA 'Transfer to Archival Storage' plattering 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 on the preservation server 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 on the network (**not** on your computer's hard drive, for security reasons), until plattering is complete and correct. Once plattering has been checked, all backup copies must be deleted, again for reasons of data security.

13. Cataloguing, indexing and DOI creation

Cataloguing and indexing of studies is a complex procedure, and as such it is usually the last part of training undertaken for Ingest processing staff.

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 Ingest team and the red folder passed back to them so they can complete the catalogue record. For those Ingest staff who have received catalogue training, the catalogue record and keyword index should be completed according to the procedures and rules detailed in the *Cataloguing Guidelines and Procedures* document.

Once the study has been plattered successfully, and the catalogue record and keyword index completed, the Ingest team will 'release' its catalogue record and create a Digital Object Identifier (DOI) number. The record will appear in the web catalogue and the data will become available for users to order.

Catalogue records are checked/proof-read prior to release, for quality control purposes. The procedures for this are outlined in the *Cataloguing Guidelines and Procedures* document.

14. Post-release procedures

Once the study catalogue record has been released, a post-release email should be sent to the data depositor. As the Discover catalogue updates overnight, the new study/edition record may not show until the following day. Therefore, it is wise not to send the email until one day after release, once the Discover catalogue record has been accessed successfully. The email contains a link to the Discover catalogue record and encourages the depositor to check the catalogue record and notify the Archive of any changes needed. A template for the release email is available. Subsequently, the email should be converted to PDF and archived.

Most 'red folders' are now electronic, but if a physical red/yellow folder exists, it should be placed in the appropriate location to be collected for storage in the Archive Safe Store.

Once the study is released, all backup copies of the study should be deleted from the processing server and individual network areas, and the pre-ingest folder should be deleted from the pre-ingest server to ensure the maintenance of good data security.

15. Checklist of main ingest processing activities

New staff may find it useful to refer to this list for each study/data collection processed.

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 (html-only) .htm format files
4. Directory structure is complete and correct, including rf directory and all files created during processing are appropriately named and archived in the correct directory
5. All electronic correspondence/data deposit/review forms/licence forms have been archived.
6. Archival storage (plattering) has been completed and checked
7. Cataloguing and keyword index have been completed
8. Discover catalogue record has been released and Digital Object Identifier (DOI) created.
9. The CALM database has been updated
10. Once the new/updated catalogue record is visible in Discover, the post-release email has been sent, converted to PDF format and archived.
11. The pre-ingest folder has been deleted from the pre-ingest server.