

External Document

Data Services Process Guides: Documentation Processing Procedures

UK Data Archive (UKDA)

Creator: Data Services

Maintained by: Data Services

Version: 01.00

Controlled document No: 78

Last Amended: 3 March 2009

Review due date: 3 March 2010

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

email: help@esds.ac.uk
Tel: +44 (0) 1206 872001
Fax: +44 (0) 1206 873001
data-archive.ac.uk

Table of Contents

1	Introduction.....	3
2	Documentation formats	3
2.1	Creating PDF documents from MS Word	3
2.1.1	Tagging	3
2.1.2	Conversion issues	4
2.2	Creating PDF documents from MS Excel	4
2.3	Creating PDF documents from text files.....	4
2.4	Creating PDF documents from other word-processing software	4
2.5	Creating Tagged Image File Format (TIFF) files from hard copy (paper) documentation.....	4
3	Creating PDF documents from TIFFs and other image files.....	5
4	Optical Character Recognition (OCR)	5
4.1	Using the Adobe Acrobat 'TouchUp Text' Tool.....	5
5	PDF documentation production standards	6
5.1	Filenames	6
5.2	General PDF editing operations	6
5.3	Amalgamating files.....	6
5.4	Cropping.....	6
5.5	Deleting/rotating pages	6
5.6	Moving pages	6
5.7	Adding notes to PDF files.....	6
6	Bookmarking.....	7
6.1	Setting bookmarks to 'Fit Page' magnification.....	7
6.2	Hierarchical bookmark structures	7
6.3	Setting the final PDF document properties	8
7	Adding headers to documentation (branding)	8
7.1	Background.....	8
7.2	Exceptions	9
7.3	Adding headers to Word documents.....	9
7.4	Adding headers to Adobe PDF documents	11
7.5	Adding headers to other documentation formats.....	12
8	Creating index files	13
9	Administrative metadata: Read and Note files	13

Document Control

Version	Notes	Last Amended
01.00	External version	2009-03-03

Replaces or supersedes:

Procedures section of *UKDA-DocumentationProcessingTechniques*

Review terms:

Annual

Is related to:

UKDA-DSS-Data Processing Procedures

UKDA-DSS-Data Processing Standards

UKDA-DSS-Processing Quick Reference

UKDA-DSS-Creating Administrative Metadata (in development)

UKDA-DSS Filenaming Conventions and Standards (in development)

Scope

What is in this guide?

This guide covers standards and procedures for the preparation and preservation of documentation for UKDA studies (datasets). For external readers, it is for information only.

What is not covered by this guide?

- Data processing – see separate document *UKDA-DSS-Data Processing Procedures*.
- Administrative metadata ('Read' and 'Note' files) that provide extra information for users and describe the processing history of each study - see separate document *UKDA-DSS-Creating Administrative Metadata* (in development).

1 Introduction

Unlike data files, there are no 'set' processing standards for documentation files. However, an ingest standard (A*, A, B or C) is allocated and recorded in the CALM processing database to denote the nature of the documentation materials deposited as part of a dataset (see separate document *UKDA-DSS-Data Processing Standards* for further details). However, value is added to dataset documentation through the creation of user guides and other documentation, and the addition of bookmarks to aid navigation.

2 Documentation formats

The primary documentation dissemination format created at the UKDA is Adobe Portable Document Format (PDF). As with data formats, successful documentation archiving requires a balance between effective archival preservation and the provision of documentation in popular and well-supported software formats to enable easy secondary use. While Adobe PDF is not an ideal archival format (though the PDF/A standard is developing fast - see <http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml>), Adobe Reader software is free and easily available on the web (see www.adobe.com). It also has advantages in that PDF documents are relatively difficult to edit, and so have some inherent protection against inadvertent change by the user.

Most documentation is currently deposited in Microsoft (MS) Office formats, i.e. Word, Rich Text Format (RTF) or Excel. These formats are relatively easy to convert to PDF.

2.1 Creating PDF documents from MS Word

MS Word files (.doc and Rich Text Format (RTF)) are easily converted to PDF documents via the Adobe Acrobat 'PDFMaker' facility. This is an inbuilt component of the Acrobat software, which may be installed directly into the MS Word interface to facilitate ease of use. Files may then be created from within Word, either via the Print menu (select 'Adobe PDF'), or using the Adobe PDF drop-down menu and selecting 'Convert to Adobe PDF'. There may also be an Adobe toolbar visible within the Word interface, with associated 'Convert to Adobe PDF' icon.

2.1.1 Tagging

Adobe PDF files can include 'tags'. Tagged PDF files improve accessibility for visually-impaired readers, as the tags provide a structured, textual representation of the PDF that is presented to screen readers. They exist for accessibility purposes only and have no visible effect on the PDF file. By default, PDFMaker adds tags to documents created from Word, but this does

increase file size. When converting large Word files, the option to turn off automatic tagging to enable easier format transfer may appear. In this case, it is best to create the PDF file without tags. The creation of tagged PDF files, though desirable, is not currently standard procedure at the UKDA; the collection contains many older datasets with PDF documentation files created from the scanned original paper (hard) copies. It is not currently feasible to tag these.

2.1.2 Conversion issues

While Word files generally convert well to PDF, there may occasionally be some problems with pagination and/or 'headers' and 'footers' falling outside the 'printable' area. These problems may be remedied by changing the dimensions of the printable area via the Print dialogue box. Extensive changes in formatting are always easiest and quickest to perform in Word rather than in the PDF file, which is less easy to edit.

2.2 Creating PDF documents from MS Excel

Excel files can also be converted to PDF files through PDF Writer, in an identical fashion to Word documents (as outlined above). However, there are some points to remember (and some pitfalls to be avoided):

- every sheet within a multiple-sheet Excel file must be converted to a separate PDF file; one conversion will only transfer the first sheet.
- it is very important to make sure all columns are set wide enough in the Excel file to display all the text within them prior to PDF conversion. If this is not done, the PDF file will display the cells with truncated text.
- large Excel spreadsheets may also cause problems in Acrobat due to the limitations of the printable area, so such conversions should be checked very carefully. Reduction to a percentage of page size is possible in Excel via the Print dialog box in order to make the sheet display on one page, but for very large Excel sheets this may render the resulting PDF at such a small size that it may need greatly magnifying to be legible.
- if the Excel file includes text or set formatting, or makes use of a 'freeze pane' scrolling facility, it may not be easy to create successful PDF conversions. In this case, the document should be left in Excel format, and a suitable note added to the Read and Note files. If the Excel file is part of the documentation for supply to secondary users, a copy should be archived with the rest of the documentation. The original Excel file should be archived with the rest of the original deposited files.

2.3 Creating PDF documents from text files

Text files can easily be converted to PDF via MS Word. Any format editing of the file should be carried out in MS Word prior to PDF conversion.

2.4 Creating PDF documents from other word-processing software

Other proprietary formats may occasionally be deposited at the UKDA (MS Works, WordPerfect, Claris Works, Open Office, etc.) If they do not include a facility for direct conversion to PDF, they may either be imported directly into MS Word, or first exported from their proprietary format as RTF and then imported into Word.

2.5 Creating Tagged Image File Format (TIFF) files from hard copy (paper) documentation

The deposit of hard copy documentation is becoming increasingly rare, but sometimes still occurs, meaning that the paper copy must be scanned to provide electronic documentation for preservation and dissemination. Before scanning, check with the depositor whether the file is available in electronic format if this has not already been done.

If the hard copy documentation is double-sided, it should for ease be photocopied to single-sided format before scanning. Material that is primarily text should normally be scanned at 300dpi resolution, though higher resolution may be used where required. All material should be scanned into TIFF format, which is a flexible and adaptable format for handling images and data within a single file. Older studies in the UKDA holdings will have one TIFF file for each documentation page; this is preferable for archival standards in case of future file corruption, and should be the norm when processing hard copy documentation, but more recent studies may have many pages in one TIFF file. The TIFF files will be further converted to PDF, but the original TIFF files must also be archived on the UKDA preservation system.

3 Creating PDF documents from TIFFs and other image files

Scanned images (TIFFs) and other image files supplied by the depositor (in any common ingest format, such as Joint Photographic Experts Group (JPEG or JPG)) are easily imported into PDF, using the 'File' drop-down menu and selecting 'Create PDF'. If any image editing needs to be carried out to enhance legibility, it is generally easier to edit the images in a graphics program such as Paint Shop Pro or Adobe Photoshop, before conversion to PDF.

4 Optical Character Recognition (OCR)

All scanned hard copy material should be subjected to OCR, except where the text content is minimal (i.e. not scans of pictures or photographs, etc.). Such TIFFs can simply be opened into Acrobat, saved as a PDF file and then inserted into any PDF document (see below for details of merging files and moving pages). Whilst OCR software is available within the current UKDA scanner, some limited OCR may also be carried out using later and 'Professional' versions of the Adobe Acrobat software (based on Adobe's proprietary Paper Capture plug-in), but this may not be available in earlier or 'Standard' versions. As licences for later versions of Adobe Professional may be expensive and limited, only certain staff within the UKDA may have it installed on their computers.

If the OCR option is available, select 'Recognize Text using OCR' and 'Start' from the 'Document' menu in Acrobat, and specify the following preferences before running the OCR:

- Primary OCR language: English (UK)
- PDF Output Style: Searchable image (exact)
- Downsample: Low (300dpi)

4.1 Using the Adobe Acrobat 'TouchUp Text' Tool

The 'TouchUp Text' tool in Adobe Acrobat may be useful for limited editing of text once OCR software has been run on scanned hard copies. After selecting the icon, place the cursor over the text to be edited. A box will appear within which the text can be amended. This tool also allows limited editing to be carried out, such as moving text along on the same line, which may be useful when pagination has gone astray. Extra lines of text cannot be added. More extensive editing will need to be done on the original TIFF file using Photoshop or PaintShop Pro.

5 PDF documentation production standards

5.1 Filenames

The UKDA has standards for the naming of PDF and other documentation files. Details of these may be found in the separate document *UKDA-DSS Filenaming Conventions and Standards* (in development).

5.2 General PDF editing operations

This section covers some of the most common procedures used in the Adobe Acrobat software.

Note: these instructions are based on Adobe Acrobat 8.0 Professional, and may differ from other versions of the software (not all UKDA staff may have access to the same version of Adobe). See the 'Help' guide and documentation for Adobe Acrobat software for alternative specifications and instructions.

5.3 Amalgamating files

From the 'File' menu in Adobe Acrobat, select 'Create PDF' then 'From multiple files'. This will show an interface which can be used to browse, select, and set the order and combination of multiple PDF files (held together in one directory).

Alternatively, open the current PDF file at the page at which another file is to be inserted, 'drag' the desired file onto the open PDF file and 'drop' it in the main text area (not the bookmark margin). Selecting the 'Document' drop-down menu, 'Insert Pages', then choosing the file to insert, will also perform the same action.

5.4 Cropping

The cropping facility is very useful for removing any unwanted marks from document pages, e.g. dark margins or staple marks from scanned hard copies. To use the cropping tool, select the cropping icon from the toolbar, or 'Crop Pages' from the 'Document' menu. Draw a box around the area to be kept; anything outside the box will be deleted.

5.5 Deleting/rotating pages

Choose 'Delete Pages' from the 'Document' drop-down menu. A dialog box will appear as a final prompt before the decision is made to delete a page. To rotate misaligned pages, select 'Rotate Pages' from the 'Document' drop-down menu.

5.6 Moving pages

To move pages around within a PDF document, click on the 'Pages' tab at the left-hand side. The images will appear in the left-hand section. Images can then be 'dragged' as necessary to a different location in the document.

5.7 Adding notes to PDF files

Notes may be added to PDF documents by selecting 'Comments' and then 'Add Sticky Note' from the 'Tools' menu. A note can be added on any page by clicking the mouse and typing in the box that appears. Notes are sometimes used to draw users' attention to anomalies in the documentation, for example filenames and formats referenced that differ from those available from UKDA. Processing staff should be aware that the 'author' of the note will appear as the licensed owner of the Adobe Acrobat software, which will often be the login name of the staff member. This should be amended to 'UKDA' by changing the 'Author' box entry under the 'General' tab in 'Properties', which may be accessed by clicking on the 'Options' tab within the

open note.

6 Bookmarking

All PDF documentation files should be 'bookmarked' to aid user navigation, whether they are in the format of a single user guide, or multiple volumes. The optimum density of bookmarking is obviously content-specific. In general, 'full' bookmarking equates to the addition of a bookmark to each smallest discrete unit of a document that is greater than one page in length. Any necessary amalgamation of files, or deletion/movement of pages within the PDF file should be carried out before bookmarks are added.

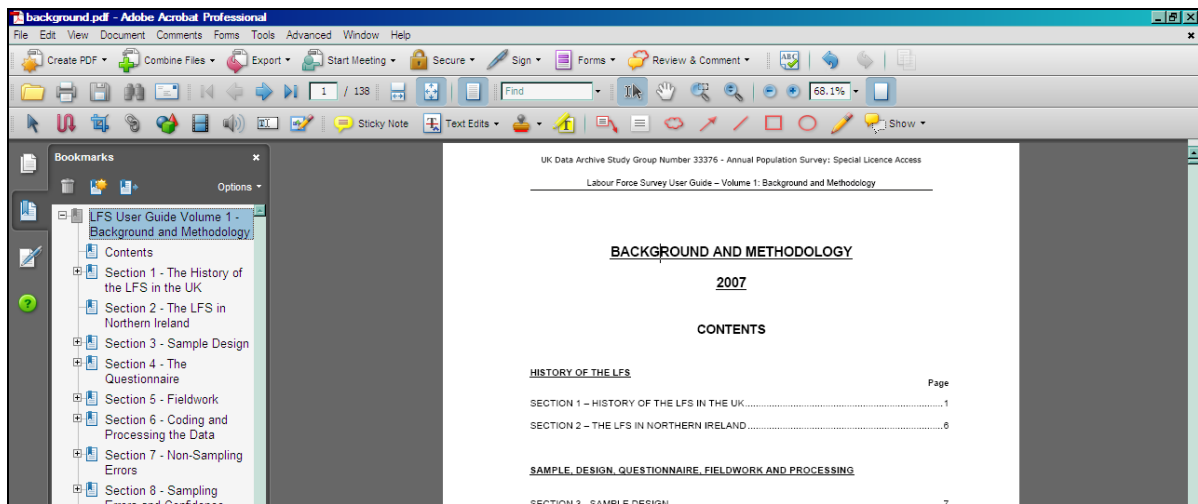
To create a bookmark in Adobe Acrobat, select the 'Bookmark' tab, along the left-hand side of the visible document screen. This will open a section at that side, in which the bookmarks will be created. Choose the 'Select' tool icon. The word(s) required on the document page (e.g. a heading) can be highlighted. Once this is done, press 'Ctrl+B', and the selected text will appear as a bookmark in the 'Bookmark' pane. Alternatively, a new bookmark may also be created using the 'Edit' drop-down menu and selecting 'Add Bookmark'. The resulting bookmark text may be edited as necessary. However, if a blank bookmark is required for text to be typed in, press 'Ctrl+B' and an 'Untitled' bookmark will appear, which may be edited as necessary. Insert the bookmark text in 'sentence case' (upper case initial letters for all nouns, lower case for conjunctions, etc.). If numerous bookmarks have been inserted, remember to save the file regularly. Many files are now deposited in PDF format, often with bookmarks already added by the depositor. In many cases, these bookmarks have been created by the Adobe 'Distiller' plug-in, and may need significant editing to conform to UKDA standards. Deleting them altogether and recreating them from scratch is often the easiest option.

6.1 Setting bookmarks to 'Fit Page' magnification

All bookmarks should be added with the document page set at 'Fit Page' magnification, which is the UKDA standard. Note that if a bookmark is added whilst the page is magnified, that is how the bookmark will be set. Bookmarks that open with pages in an assortment of magnifications make for a very unprofessional appearance, and should be avoided. Files may arrive from the depositor with bookmarks set like this, in which case all bookmarks must be reset to 'Fit Page'. Remember to proof-read bookmarks thoroughly; there is no facility within Adobe Acrobat to spell-check them.

6.2 Hierarchical bookmark structures

Bookmarks should be 'nested' within a hierarchical structure. As a general rule, if the document contains a comprehensive contents page, the titles of the bookmarks and their hierarchical structure should reflect that. When all bookmarks have been created, the hierarchy should be closed leaving just the top bookmark visible (normally 'User Guide'). However, there are some special cases, for example the datasets deposited by the Centre for Longitudinal Studies, where the bookmark hierarchy must be left extended with all major bookmarks visible.



Example of a hierarchical bookmark structure

6.3 Setting the final PDF document properties

Document properties should be set so that the PDF document opens with the bookmarks panel visible, and so that the correct document metadata (now routinely read by internet search engines such as Google) settings are present.

To do this, go to the 'File' menu and select 'Document Properties' and add the following settings:

Under the 'Description' tab, add a suitable document title in the 'Title' box, such as 'General Household Survey Questionnaire 2005'. Under 'Author', add 'ESDS' or 'UKDA' as appropriate, or if the PDF file has been created by the depositor prior to deposit, the original organisation name should be used. Current practice is to leave the subject field blank, though this may change as metadata becomes increasingly important in the future.

Under the 'Initial View' tab, to ensure that the document opens with the correct magnification and that bookmarks are displayed, the following items should be checked:

- Show: 'Bookmarks and Page'
- Page Layout: 'Single Page'
- Magnification: 'Fit Page'
- Window options: 'Centre Window on Screen'.

7 Adding headers to documentation (branding)

7.1 Background

Google and similar web searches now return results that include searchable PDF documents. If a search results in a 'hit' on a document on the UKDA/ESDS web site, it may not be obvious to the searcher that the document is part of a dataset held at the UKDA/ESDS.

Therefore, where possible, UKDA will use an informative header to 'brand' the first page of each PDF file created as part of dataset documentation. (The principle is similar to the addition of a header to qualitative interview transcripts.) This policy applies whether the documentation consists of one file or multiples. Note that the presence of a UKDA header does not imply any claim on copyright, which remains with the original document copyright holder. It is merely a branding device used to aid web searchers and identify a component of the UKDA collection.

7.2 Exceptions

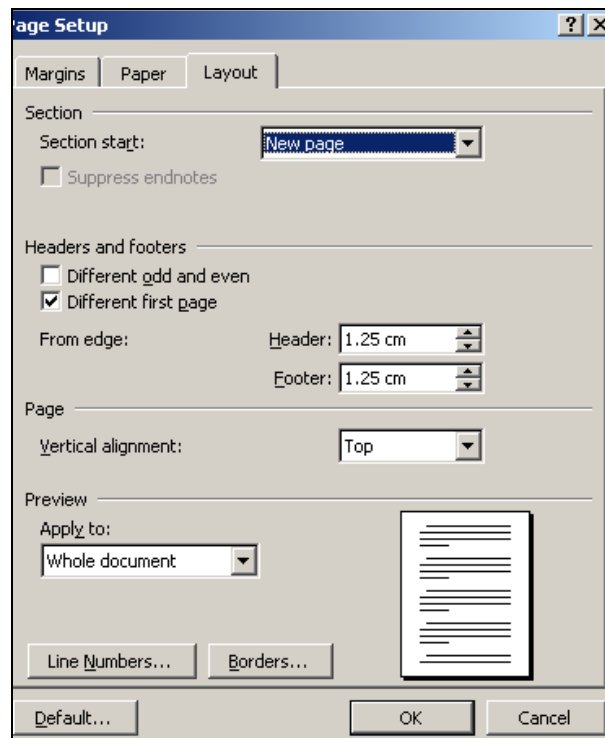
There are some occasions where branding with suitable headers may not be possible or desirable.

- **Documentation received from the Centre for Longitudinal Studies (CLS).**
- **'Locked' (password-protected) PDF files where no editing is possible.** Some PDF files may have been created with a high level of security, which limits user options. Unless the password to 'unlock' them is available, neither headers nor bookmarks can be added. This may occur when, for example, the depositor is a government department and the PDF file has been created by a survey contractor. All reasonable efforts should be made by the UKDA to obtain either a Word version of the PDF file, an 'unlocked' copy of the PDF file, or the password. If this proves fruitless, it should be recorded in the study Note and Read files as appropriate that the file was unable to be edited and so it has not been processed to the usual UKDA standard.
- **The depositor objects to the presence of a UKDA header in their documents.** Most depositors will not mind a discreet header, but if a request to remove it is received, the files should be edited accordingly and replattered. The depositor's objection should be recorded in the Note file for future reference, and the Data Services Manager informed (who may take over liaison with the depositor as necessary).
- **Software used for documentation file creation limits or precludes header creation.**
E.g. Windows Help (.hlp) files.

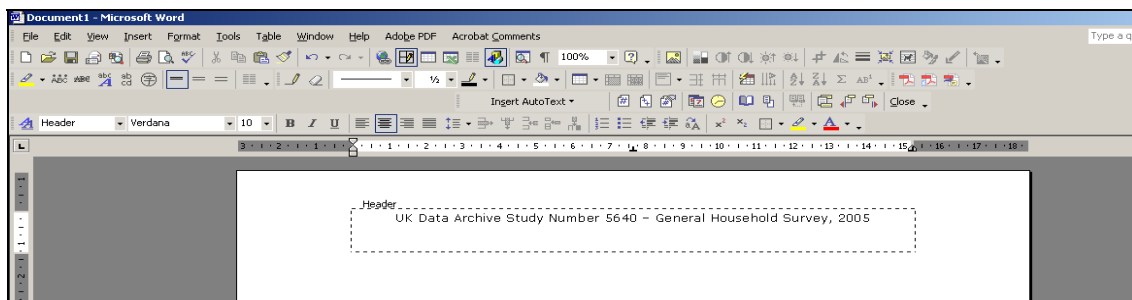
7.3 Adding headers to Word documents

Where documentation is received in Word or RTF, the addition of a header may be done relatively easily. First, make a copy of the original to work on (the depositor's original should remain as received without header). The copy file will need to be saved once the header is added before it can be converted to PDF, which is another reason for working on a copy rather than the original. If under any circumstances the original is used, ensure that the header is removed from it once the resulting PDF file has been checked.

- Open the document in Word.
- Find 'View' on the top menu and select 'Header and Footer'. A 'header' box will open, where text can be added.
- To add the header to the first page only, on the 'Header and Footer' toolbar, click Page Setup (the 'open book' icon), and the 'Page Setup' box will open:



Select the 'Layout' tab, then tick the 'Different first page' box, and then click 'OK'. This will ensure that the header appears on the first page of the document only. (Further information may be found in the help guide for the version of MS Word in use.)



Add the study number and title in the following format:

UK Data Archive Study Number 5640 – General Household Survey, 2005

Style specifications:

- Font: Verdana, normal (i.e. no bold or italics).
- Size: 8pt (large enough to be legible and small enough to be unobtrusive).
- Justification: centred, and at the top of the header box, so that it is suitably distinct from the text in the body of the document.
- Ensure that there is a spaced hyphen between the study number and the study title.

This style has been agreed as a UKDA standard, and should be used. However, it is possible that this style may cause display problems depending on the document. For example, if the header text appears too near the first line of the document text, the page margins can be adjusted (see 'Page Setup' box above) to move it further away.

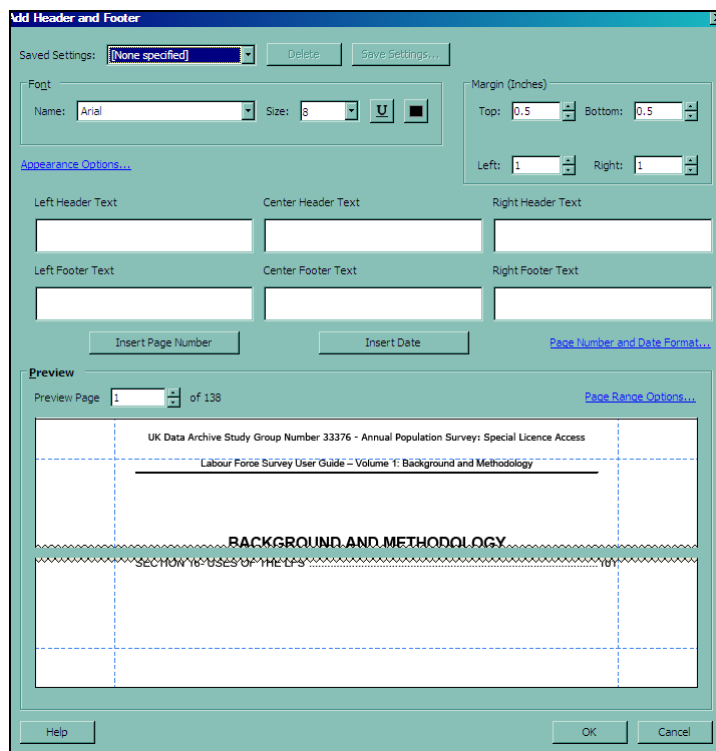
When the header has been added, save the document copy, and make any other adjustments deemed necessary before creating the PDF file according to normal documentation procedures. The PDF copy must be checked to ensure that header transfer has been successful. If so, the Word file copy may be deleted.

7.4 Adding headers to Adobe PDF documents

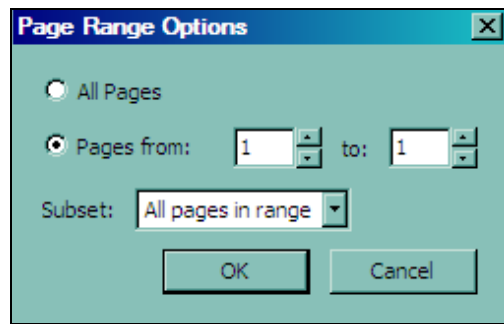
Where no Word or RTF copy is available, the header may be added within the PDF document.

Note: these instructions are based on Adobe Acrobat version 8 Professional, where the interface differs considerably from previous versions (It is possible to add headers in versions 6.0 Standard onwards). The procedure is as follows:

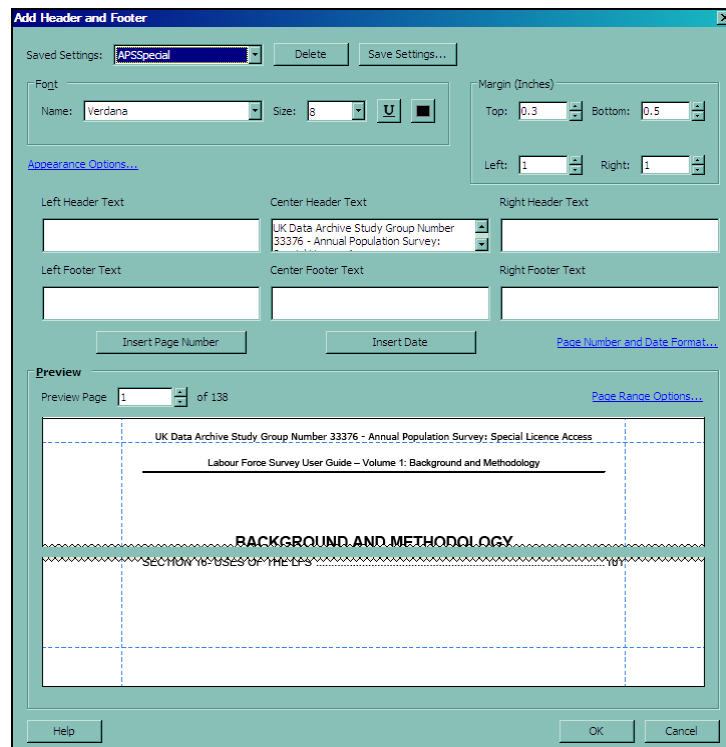
1. Open the PDF file.
2. Go to 'Document' on the top menu, then choose 'Add Headers and Footers', then (usually) 'Add' from the selection of 'Add', 'Update' or 'Remove'. If there is an existing header or footer, a further prompt will then appear to 'Add new' or 'Replace existing'. Usually, 'Add new' will be the option of choice. Once this is selected, the following box will appear:



3. Select the following font settings:
Font: Verdana, size 8.
Margin (Top): 0.5 (may be adjusted to 0.2 or 0.1 if another header is present).
4. Add the required header text into the 'Center Header Text' box to ensure central justification. The text wrapping does not matter here, as it will appear in one line on the page.
5. Click on the 'Page Range Options' hyperlink, select 'Pages from', and adjust the range to show 1 to 1 (see below), so the header will appear only on the first page. **Note that 'All Pages' is the default option - if this is not reset, the header will appear on every page and will need to be removed (using 'Add Headers and Footers' > 'Remove' from the 'Document' menu), and then the process started again to add the correct header to the first page.**



6. Click 'OK' and then view the header as it appears on the first page. If alterations need to be made, repeat the process. The position on the page can be adjusted within the 'Margin' (Top) section as necessary.



7. Once the header is correct and added to the first page, click 'OK', check the display in the Adobe PDF window, save the file and repeat as necessary for each document in the dataset, and complete processing as normal.

Note: it is possible to save settings in Adobe Acrobat based on a basic header creation, for addition to subsequent documents. Obviously, the study number and title will need to be edited according to the study, but settings such as header on the first page only, and font and font size can be saved so that they do not need to be reset on each occasion.

7.5 Adding headers to other documentation formats

Excel

Header information will only display in printed Excel files, not in the normal spreadsheet view, which rather defeats the object of adding a header for these purposes. As it is currently unlikely that a Google web search will pick up an Excel file, adding a header should depend on the nature of the Excel file. If it is possible to identify the file as part of UKDA documentation in some form, it should be done. For an example of how to do this, see the Excel file 5640_changes_2004_to_2005.xls included in the GHS 2005 documentation.

Powerpoint

Documentation may occasionally include Powerpoint presentations, for example, the Family Resources Survey (FRS). It may be possible to create PDF files from Powerpoint, but the same principles should be applied to these as to the Excel files, i.e. header addition should depend on the nature and contents of the file.

HTML

It may not be possible to add header information to .html documentation pages, but again this depends on the nature of the file. Procedures for dealing with .html documentation are currently in development.

Software package files (e.g Windows .hlp)

Header addition is unlikely to be possible.

8 Creating index files

Some older datasets in the UKDA collection with multiple documentation volumes may have had a hyperlinked 'Index' file to the documentation created. This is no longer necessary now that the UKDA has moved to a more intuitive filenames convention for documentation. It is also not desirable if the UKDA is to move to the PDF/A documentation preservation standard, as hyperlinks to external files may not be permissible. However, a certain degree of flexibility is permitted so that an index file may be created for those studies with a lot of documentation in multiple formats (e.g. the *Family Resources Survey*). If it is felt necessary to create an index file for a new study or series, please consult a senior member of the Data Services team for advice.

9 Administrative metadata: Read and Note files

In addition to the documentation supplied by the depositor and processed as described above, additional documentation (metadata) is also created by the UKDA. This consists of the html format 'Read' and 'Note' files, created via the CALM database to accompany each dataset processed at the UKDA. Procedures for creating these files are covered by a separate document, *UKDA-DSS-Creating Administrative Metadata* (in development).