Steps to Digitising Church Records and Documents

By David Parker ©
22 October 2016 v 1.0
For feedback and further information, contact archives@qb.org.au

The Digitising Project

Digitising documents is a good way to store them, make them widely accessible to users, save space and to conserve them. However, the job must be done to appropriate standards or else time will be wasted, the results disappointing and the records and their information put in jeopardy. Although there are many benefits to digitisation of existing documents, it must be noted that the process can be time consuming, and costly; it requires specialised equipment, software and training. Success will also depend on the quantity of documents involved.

The aim of any archival digitisation project is to produce good clear machine readable (OCR) documents that are bookmarked and indexed, which are of such quality that they can be used in place of the originals both on-screen and as print-outs.

For a general introduction and some overall considerations, see our guide, Digitising Church Records – A Guide for overall matters (http://www.qb.com.au/wp-content/uploads/2014/10/Archives-Resource-Digitising-your-Church-Records.pdf) and also, Guide to Digital Archiving http://www.qb.com.au/wp-content/uploads/2014/10/Archives-Resource-Digital-Archiving.pdf

The basic choice is for digitising to be out sourced to a commercial operator, or for it to be done in house. Baptist Church Archives can make suggestions about commercial operators. The advantages of outsourcing are good quality results with little effort, and it obviates the need to set up an in-house system thus saving on equipment and training costs. These advantages are especially relevant where a finite batch of material is to be scanned; this is the situation for most churches where there is a batch of documents from the origins of the church to recent times, (say 2000-2010, when digital recording commenced.

This document describes the in-house process using Adobe Acrobat DC (Adobe Acrobat Pro) and Abbyy Fineview (optional) software. Note that the free application, Adobe Reader (and equivalents from other sources) is not sufficient for this process, but can be used to read the final products. Recent earlier versions of Adobe Acrobat will probably be suitable. For more information, see your IT retailer or htttp://www.adobe.com

See also Pixedit http://www.pixedit.com/ which is fully featured and designed for high volume automated operations. It is very expensive.

There are many different types of PDF software on the market (eg Nitro Pro) but most do not seem to support the functions needed for this process. So Adobe is probably the only option.

Note that Adobe Acrobat is available as a stand-alone downloadable version, or an on-line subscription version (internet connection needed). Both versions operate the same and have the same functions, but the difference is a one-off cost with periodic extra cost for updates in the case of the stand-alone option, over against regular monthly payments, always having access to the current version and the need for internet access.

The subscription version would be suitable where a finite batch of material can be processed over the period of a few weeks after which the subscription can be cancelled. The total financial outlay for the project would probably much less than the cost of the stand-alone version.

(Note that Adobe has an educational licence option which is considerably cheaper.)

Computer system

PC/Laptop

Any modern PC or mid- to high-range laptop, equipped with word processing, spreadsheet and other relevant software, plus the above, will be sufficient for scanning and processing of documents. However, for storage, access and use of the files, and all the other electronic files used in the church's administration, further infrastructure would be needed.

In addition, for awkward, large documents which cannot be placed on a flat bed scanner or for fragile ones that cannot be physically manipulated for scanner copying, a camera copying system would be needed. This would ideally require a camera capable of capturing in RAW format. (Normal domestic cameras only offer JPG format. This is a compressed format, which loses some definition initially and further deterioration each time the file is saved.) RAW images are converted to TIFF which become the master records, from which working copies in JPG etc are made.

Scanner

Condition of original paper documents and the text/images to be scanned will affect the type of scanner to be used. If only relatively new separate sheets of paper in good order (typically A4 office style paper) with clear text and images are to be scanned, then a copier with an automatic document feeder (ADF) can be used (such as are common on office photocopiers and some multifunction units). Modern units can scan single or double sided to a variety of formats and resolutions, at a fast speed. However, it is necessary to check the quality of the resulting scans to ensure that they meet the standard required. A dedicated flat bed scanner is likely to produce better quality.

However, in many cases, the documents to be scanned will likely be older, possibly fragile or damaged, in different sizes, and with faint handwriting, or they may be photocopies (even spirit duplicated), or computer printouts from a dot matrix or light sensitive paper. Often the text will be in poor, uneven or faded characters. In these cases a flat-bed scanner will be needed.

Commonly available flat bed scanners will only handle A4 size, and are likely to be too small, so a larger dedicated A3 size scanner will be needed. Most record books (minutes financial) are larger than A4. They are also likely to have problems opening up flat because they have many pages and tight binding, which will make it difficult to get a good clear scan and to reach the gutter edges of the page. So a 'book scanner' is needed, which is designed to allow the book to be opened partially and for the scanning bar to reach the inner edge. (Eg, Plustek A300 http://plustek.com/usa/products/opticbook-series/opticbook-a300/)

If slides and film negatives are to be scanned, a special film scanner will be needed. Film scanner adaptors for home A4 flat bed scanners do not produce sufficiently high resolution.

Policy regarding original documents

Digitising of paper documents requires a decision about the disposal of the original documents after scanning. The choices are:

1. Retain the paper documents for reference, but use the digital versions as back up and for sharing, researching and for convenience etc
2. Retain the paper documents but place them in deep storage for permanent retention, relying on the digital versions for all uses except in the most extraordinary circumstances
3. Destroy the paper documents (in the interests of saving storage costs etc), relying entirely on the digital versions. There may be a time delay in the destruction process to enable sufficient checks that the digital system is working as expected.

Standards

Digital version should be stored in a file format that is likely to be readable as widely as possible by different computers, and as far into the future as possible, while retaining the 'look and feel' of the originals. The current standard for this is PDF/A.
Documents should be scanned in colour where colour is an important part of the original and/or where there is likely to be a legibility problem (eg with old, discoloured paper or ink). Scanning should be at 300 DPI although if storage space and computing power are no problem, 600 DPI can be used. A lower DPI can be used for documents consisting of purely black and white text or images.
Photographs and other images - 600 DPI where possible.
Note that if Option 1 under Policy is chosen, a lower resolution may be acceptable.

Scans should be made to TIF format in high resolution. This is a lossless format, and is retained as the master mint copy from which any working copies are made. The next step is to produce working copies in PDF for documents and JPG for images, and then to process the PDFs for regular use.

Preparation for scanning

Check condition of paper and binding, taking great care in physical manipulation when fragile, and with ADF in the case of separate sheets.
If the paper or ink is badly discoloured, adjustment of scanning parameters may be needed to obtain a satisfactory image.
If binding is loose and badly damaged it may be better to de-bind, scan as separate sheets and re-bind.
Make sure all pins, staples, scotch tape etc are removed.
Adjust software for the relevant page size. A white paper mask around the paper may be helpful to avoid need for excessive electronic cropping.
Set the scanner to save in TIFF format.

Scanning

When scanning manually, be sure to be consistent in placing paper on the platen in the same position and properly square to minimise the need later for cropping of excess space, and adjustment of the image (straightening etc).

For books, scanning a single page only at a time will simplify the processing of scans later, but is more time consuming and likely to be more physically damaging to the original because of each page has to be treated separately, requiring more physical handling of the book.

An alternative is to place the book flat down on the platen and scan two pages at a time. This is only feasible if (i) the binding is flexible enough to go flat without damaging the book, and also (ii) if the scan of the two pages can be split later into single pages.

AbbyFine print software supports splitting of facing pages into a single pages. Select all TIFF images in the batch, right click, and select convert to PDF in Abbyy. Double pages are automatically converted to single pages in the right order in the resultant PDF.

Note that this process of dividing double pages is possible with Adobe Acrobat but it is a complex process. It involves cropping pages to eliminate one half, and printing to a new PDF; then cropping the original file again to eliminate the other half and printing to PDF; then final reassembly of the 2 half-page files in the right page order! See on-line for some instructional videos and web pages with detailed instructions.

Some scanners come with software which produce PDFs machine readable (or searchable, or OCR – optical character recognition). If this process is not done on creation, it can be performed in Adobe Acrobat (see below).

Editing of PDF in Adobe Acrobat

Once the scans have been converted from TIFF into PDF, it is time to start processing them. Use Adobe Acrobat for this process.

Adobe Acrobat supports these functions which are relevant to processing of scanned pages:

- merging and splitting files;
- deleting or adding individual or groups of pages, including adding a cover sheet;
- moving pages into required order;
- optimising the file to reduce size, improve look of page etc;
- cropping pages to eliminate unnecessary parts of the page which may have binding marks, edges of the pages, dark shadows etc (cropping can be done page by page or in batches);
- rotating pages; annotating pages, adding comments, sticky notes etc
- indexing (a single document or group of documents);
- bookmarking of pages and headings for ease of navigation by users;
- setting defaults for when document is first opened (eg whether it is whole page view, displaying bookmarks etc);
- security (eg controlling if document can be printed, edited, copied, documented, disassembled etc), encryption
- saving the file in different PDF formats (especially PDF/A)

Familiarity with the basic operation, functions and features of Acrobat is assumed. (There are many source of help and instruction on-line and via the Adobe support system to assist this process.)

A suggested workflow

It is suggested that the file be saved at the end of each of the steps below with a different file name (eg, add a number in sequence – 1, 2, 3 or a suffix indicating the stage reached – eg, edit, opt, def, ndx,) so that, in the event of problems, a previous safe version can be accessed, thus avoiding the lost of work. Rename the working file to the final title at the point mentioned below (see *). Delete temporary files after full checks have been done at the end.

1. Check the PDF that all the pages are included and in the right order – Select - organise pages
2. Add cover sheet if required – create a new PDF page using word processor (save to PDF) and add this page to file: Select Organise Pages – Insert, or use Combine files function
3. Crop pages to remove any excess white space, shadows/marks from gutter, page edges or other marks (either page by page or in a batch) Edit PDF – Crop pages
4. Edit pages to remove any unwanted marks (eg discoloured, annotations, etc)

(Note that this Adobe Acrobat supports other advanced editing functions as well, but these are not covered here.)

5. Bookmark pages for easy navigation by user (title, contents, chapter, sections etc) as desired: Control B or Edit – More – Add Bookmark
6. Setting default opening view of file by user:
    a. Select File – Properties, Description – insert title, author, subject, keywords as required
    b. Select File – Properties, Initial view - select options from Navigation Tab, Page Layout, Show and any other options
7. Optimize file, including reducing file size if very large
    a. Tools – Optimize – Optimize Scanned Pages – select optimization options and also Text Recognition options as required
    b. Tools – Optimize – Reduce file size

8. *Check that the file is as intended and then rename file to its final title here.
9. Save file: Tools – PDF Standards – Save as PDF/A
10. File can be indexed for fast search (this is much faster than the normal word search of an OCR file using Control F). The command for fast search is Control – Shift - F
11. For embedded index: Select Index – Manage Embedded Index – embed index
    a. To index a batch of files, place them all in the same folder with no other files. Select Index – Full text index with catalogue and then complete the boxes on screen
    b. Note that this will produce a .PDX file in the same folder, and a new sub-folder with 2 .IDX files. All these files (original PDFs, PDX, and the sub-folder must all be present. To share the PDFs with catalog index, copy all)
12. To impose security restrictions on the PDF (eg limiting who can read or access it by password, restricting printing, copying etc) Select File – Properties, Security
13. After final check, delete temporary files.

Photographs

The purpose of digitising photographs will determine the process. There are two main options (cf 'Policy on original documents' above):

First, to use the digital versions as working copies while retaining the originals in the church's archival collection, or second, to dispose of the originals or at least, place them in deep storage with only rare use.

In the first case, the digital scans can be of low resolution and the identification can be basic because the original paper photographs can easily be accessed for any serious use, including displays. Further scans can be taken if any work is needed to the scan to offset the effects of damage to the original photograph.

In the second case, the scan must be of the highest resolution in a lostless format (TIFF) which is saved as the mint copy, from which working copies in JPG or other format allowing for small file size. Also, any attempts at restoring damage to the photograph (caused by discoloration of paper, bends, scratches, insects) should only be carried out on a copy, not on the original mint copy which should be as close as possible to an exact replica of the original. Because the original photograph will not be accessible (or no longer existent), the index of the photographs needs to be completely comprehensive including a full physical description of the photograph (size, condition etc) as well as recording, where known, subjects, occasion, photographer, location, date, etc.

Photographs should be removed from frames for scanning, although careful use of a camera may produce adequate results.

Large photographs may not fit on a flatbed scanner and will need to photographed.

END