Conversion of printed club journals to electronic storage, and specifications for electronic publishing

Bruce Welch

Sydney University Speleological Society
bruce@bookproduction.org

Choice of Scanner

Most scanners would be suitable for scanning club records, this includes journals and photographs. The problems arise with ensuring that you scan once so that the scans can be used for a range of purposes. There is little point in scanning journals without the resolution and format being suitable for Optical Character Recognition (OCR). Similarly, it is a wast of time scanning photographs and slides at low resolution for web viewing; it is much better to scan initially at high resolution and then, if necessary, make copies of the scans and reduce the resolution for web etc.

If journals or books are bound it is often difficult to get a good result without destroying the journal or book. I have a Plustek OpticBook A300 which is an A3 scanner where the glass and scanning head goes right to the edge along one side. This allows a book or journal to be scanned without breaking the spine.

The optics and software on this scanner allow for superb results. The usual method is to scan to a multiple-page PDF which has a high resolution (600ppi) JPG image behind each page. These images can be edited by opening the PDF file in Adobe Acrobat Pro and then opening the page within Adobe Photoshop. The normal editing is usually to clean up any stray marks and perhaps to repair some broken type.

Loose-leaf scanning is easier, especially if you have access to an automatic-feed double-sided scanner. Don't forget that for good OCR the scanning must be at 600ppi.

There are scanning systems which use one or two digital cameras to photograph each page of a document, however the quality of the scan is sometimes not suitable for OCR.

Reflective scanning is used for books, journals, photographic prints etc, while transmission scanning is used for slides and glass plates.

In the end, you get what you pay for in a scanner. An expensive scanner (mine cost $2000) has good optics, good software and good hardware.

To scan a book or journal which contains photographs, the images will have to be scanned separately (either as greyscale or colour) at 300ppi, descreened in Photoshop (or similar) and then placed back into the PDF using Adobe Acrobat Professional.

Again, Acrobat has the best OCR software, other programs may be OK but I cannot vouch for any of them. Sometimes pages are rotated (from portrait to landscape) during the OCR process, in which case rotate them back again and save the file.

Storage & Hosting

Once you have scanned your valuable records make sure that you store them on a number of different mediums and in different locations. Storage options range from CDs & DVDs, external hard drives and hosting on fileservers. Remember that CDs & DVDs are not archival and external hard drives can fail. Solid state storage is becoming cheaper but there are still reliability issues. The Linnean Society of NSW is now having all its back issues hosted on the University of Sydney Library eScholarship website (back issues are at http://ojs-prod.library.usyd.edu.au/index.php/LIN/issue/archive, the current issue is at http://ojs-prod.library.usyd.edu.au/index.php/LIN/index and all the different organisations are at http://ojs-prod.library.usyd.edu.au/). I consider this to be one of the best options for permanent storage of your files as they are properly backed up.

You can have the files also on your own club web site and you can create a CD/DVD for distribution to members. The ASF is doing a DVD of all the Conference Proceedings which will be available soon. This DVD will also have on it a combined contents of all the proceedings so that this one file can be

searched quickly for any keyword(s) such as the author's name, subject or even just a cave name. Naturally it is possible search every PDF file on the DVD using search "All PDF Documents in" and select the appropriate folder.

Do not throw away the journals once you have scanned them. If you don't want to store them make sure they go to a library such as the National Library, Mitchell Library, State Library of Victoria etc. Give them a copy of the digital files too.

Converting files to HTML is not really a good idea as it is hard to extract the text and all the images are drastically down-sampled. ePub files are OK for novels but I consider them a waste of time for most caving related files. High-resolution PDFs, TIFFs and Word are likely to be readable for a very long time.

Scanning Images for Archiving

The museum standard for photographs is 300ppi at A4 size (or 10" x 8" if that is what the print size is). Postcards should be scanned at 300ppi at 200% size (or 600ppi at same-size); smaller prints should be scanned at 300ppi at 400% size (or 1200ppi at same-size); slides should be scanned at 300ppi at 800% size (or 2400ppi at same-size) however in all cases scan at higher resolution if you can. Scan all photographs in colour even if they are "black & white" as this will better capture the full tonal range.

Save all images as TIFF files as this is a lossless format. If required you can save a copy at a lower resolution as a JPG file for emailing. Unless absolutely necessary, do not use your scanner software to manipulate the file as scanner software is often very poor, it is always best to do this in a proper program such as Photoshop.

Storage of your files is the same as for your journals although Flickr is very good and a lot of libraries and museums use it. I have put up about 300 images for the Jenolan Caves Historical & Preservation Society – see
http://www.flickr.com/photos/jenolan_history/

Microsoft Word Rules

Ignore these Word rules at your peril!

- Don't do double returns between paragraphs
- Don't double-space after sentences
- Don't use multiple tabs to space or align
- Don't use multiple columns
- Don't use manual page breaks
- Don't place photos or graphs in Word
- Don't try complex formatting
- Don't make a number of files into one file
- Use paragraph styles if you know how
- Turn Paragraph marks on
- Word set-up for PDF creation
- Do not use Printer Metrics to lay out your document
- Do not allow A4/Letter resizing
- Embed ALL fonts
- Check that there is NO font substitution

Which software to use for creating a PDF from Word

My preferred program is PDF Creator which is a free program. It is important to change the <Options><PDF> to "Prepress". When installed, PDF Creator sets up a "virtual printer" in with your other printers, this also happens when you install Adobe Acrobat Professional. In Word (or any other program) go to <File><Print...> and select one of these virtual printers. If you use the Adobe PDF printer, make sure that you select "Press Quality". With these setting both will create a file which will not adversely down-sample the file and will ensure that all fonts are embedded.

Once you have created your PDF file it is very important to check it. The best way to do this is to print it out and make sure that nothing has changed from your original Word file.

Specifications for Electronic Publishing

Always supply images as TIFF files. If for some reason this is not possible, make sure that when the file is saved as a JPG that you select "Quality 12 Maximum large file" otherwise the file will be degraded. Images must be 300ppi (or larger) at the size that they are to be reproduced. Never supply images in a Word document as this may reduce their resolution and these images are very difficult to extract from a Word file. Similarly, supply Excel files separately, either as xls/xlsx or as a PDF (created using PDF Creator or Adobe Acrobat).

Line drawings and line art should be scanned at 1200ppi (minimum 600ppi) but not as greyscale or colour; your scanner will use terms such as line, text, black & white, bitmap. Vector files (ie files created in Corel Draw, Adobe Illustrator etc) should be supplied as EPS or PDF files with the fonts embedded.

Remember an original is NOT . . .

- A print from a slide
- Colour or black & white copies of a print, drawing or previously offset printed image
- Colour or black & white print-out of a scan
- An image taken from the web
- A file from your camera that has been opened & resaved on your computer as a JPG, and/or emailed

An original is . . .

- A first generation slide,
- a print (from a negative) or a drawing
- A file transferred from your digital camera (and not opened & resaved)
- A previously offset printed image

PPI; LPI; DPI; SPI

dpi* = dots per inch* ~ (e.g. scan @ 300 dpi)

spi = samples (spots) per inch (e.g. scan @ 300 dpi)

ppi = pixels per inch ~ held in computer

lpi = lines per inch ~ screen ruling applied

dpi = dots per inch ~ output imagesetter or digital printer

* some scanning software refers to dpi when they really mean spi and/or ppi

Some Tips for the use of Adobe Programs

InDesign

Life will be a lot easier for you if you customise the Preferences to show “Custom Tracking/Kerning” <Edit><Preferences><Composition>. You should also “Show Hidden Characters” <Type><Show Hidden Characters>. This will allow you to see how many spaces there are between each word, where the tab marks are, and where the paragraph marks are. This is particularly important in troubleshooting text that has been imported from Word. You also might like to colourise your menus <Edit><Menus...>.

Use paragraph styles and if you wish to import a multi-page PDF use the freely available script. Be careful with colour, it is important that all swatches are CMYK and that there are no RGB swatches. To simplify your file remove all unused swatches except C,M,Y & K. Set up a custom “Preflight Profile” so that InDesign is constantly checking for low resolution images, RGB colourspace, etc.

One good tip if you are having trouble with Word text is to copy and paste it into Notepad, then copy and paste it into InDesign. This will strip out all hidden formatting, however be careful as this method will not retain italics. In fact pasting any text into InDesign will behave in the same way. Place all images and text using <File><Place> (Ctrl D), never paste or drag images as this will cause problems further down the track.

Once you have finished your job in InDesign package the file <File><Package>, then close the current InDesign file. The reason for packaging the file is that by so doing, InDesign will check all the links, create a new folder called “Links” and collect all the linked files into this folder. InDesign will also create a folder called “Fonts” and collect a copy of all the fonts used in your document into this folder. The new master folder that InDesign has created (which contains a copy of the InDesign file, plus the Links folder, plus the Fonts folder) is the one that you should archive.

To create your high-resolution PDF file navigate to the new InDesign file and open it. Then go to <File><Export> and make sure that the “Save as type” is set to “Adobe PDF (print)”. Use the “Adobe

PDF Preset” called [Press Quality] and under “Marks and Bleeds” tick “Crop Marks”, set the “Offset” to 5mm and tick “Use Document Bleed Settings” which should be 5mm. Once you have created your PDF file pre-flight it (see the Acrobat section of this paper for details).

Photoshop

The use of Photoshop is enhanced with the use of menu colours, go to <Edit><Menus...> and select the colourisations you like for the frequently used or difficult to remember menu items, for example I always colourise “Image – Mode - CMYK”. One semi-hidden item is the expanded options in “Shadows/Highlights” <Image><Adjustments><Shadows/Highlights>. Make sure the “Show More Options” box is ticked, then set the “Shadows” to 0-30-30; the “Highlights” to 0-30-30; the “Adjustments” to 0-0 and then click “Save as Defaults” By doing this you will ensure that every time you use this tool it will open without any adjustment to your image, and by sliding the Shadow-Amount; Highlight-Amount and the Midtone Contrast (with “Preview” ticked) you will see the effect of these adjustments.

If you are new to Photoshop CS6 you will like the new straightening tool; this is available when you have the Crop tool selected.

Most scans and many photographs require a little “Unsharp Mask”. To de-screen an image that has been scanned from a previously printed image, use “Gaussian Blur” and then “Unsharp Mask”.

Acrobat

Unfortunately Acrobat X (contained in CS6) hides some of the previously easy-to-find menu items. The functions to reduce the size of a PDF, export to Word, export to postscript etc now appear under <File><Save As...>. The ability to combine a number of PDF files into a single PDF is now most easily accessed by closing the active PDF document (but not the program) and using the menu item which appears on the screen. I also find the functionality of Acrobat is greatly enhanced by customising the menu bar at the top of the screen. To this I add Rotate, Insert from File, Crop, Edit Document Text, Edit Object, In This File (OCR), Encrypt, Export All Images, Output Preview, Preflight, and Advanced Search.

You can use the default “Preflight” profiles, however you should set up your own, particularly for mono (black only) work. Most printing companies will supply you with their own preflight profile(s) which you can import straight into Acrobat.

Once you have created your PDF file using Acrobat you must check it. The first thing to do is to “Preflight” the file (using the button you put on the Menu Bar). If Preflight has been set up properly it will tell you about RGB in colour and mono jobs, more than one colour in a mono job, missing fonts, images and line art below the minimum standard, etc. You can also check the “Output Preview” to have a look at the colour separations. Once all this is done, make any corrections and re-PDF the file if required, and then print the document out and check it carefully.

Sometimes a problem PDF file that you have received from someone else can be fixed by saving it as a postscript file (.ps) and then recreated with Acrobat Distiller (using the correct settings of course).

Remember also that PDF files can be opened and edited in Illustrator.

Colour and Screens

The human eye can see a much wider range of colours than can be displayed on a computer screen (which uses RGB) – this range is called colour gamut. The process printing (CMYK) represents an even smaller range of colours.

RGB stands for Red Green Blue and when used on a computer screen full intensities of Red, Green and Blue give white – this is called additive colour. The opposite is true with CMYK (Cyan Magenta Yellow Black) where full intensities of CMY give black – this is called subtractive colour. In colour theory just CMY would be a true black, however technical limitations mean that a black printing plate is also added to give depth to the colour.

Because your computer screen can only give an approximation of the actual colours which are in your file (and also because most desktop printers also cannot properly represent the file) it is important to calibrate your monitor. The method of doing this depends on the operating system and the type of monitor you have. It is a common trap for people to alter colours on an uncalibrated monitor and then to be very disappointed when the final book comes back from the printer. This problem applies to mono photographs also. If in doubt get a commercial proof done.

Moiré

When a previously printed image is scanned and used in another publication a moiré pattern will appear. Unless you correct this before the job goes to the printer, you will have no idea how bad it might be. It is caused because the halftone screen applied by your printer clashes with the screen in the original image, how bad this clash is depends on the screen ruling and the screen angle of both. To get around this problem you must descreen the image – this is done in Photoshop by applying Gaussian Blur (at a level where the dot stops being visible) and then applying Unsharp Mask to bring the image back to look fairly close in sharpness to the what you started out with. Again how much you have to apply of each of these depends on the screen ruling (number of dots per inch) of the original.

Bruce Welch
21 Thompson Street, Marrickville NSW 2204 02 9569 9928
bruce@bookproduction.org