UNIVERSITY OF Southampton

Report

Title: **19th Century Pamphlets Online Scanning Guidelines**

From: BOPCRIS, Hartley Library, University of Southampton, Southampton, SO17 1BJ

Tel 023 80598730

Email bopcris@soton.ac.uk

Date: March 2009

Contents

Introduction

The following guidelines were used by BOPCRIS within the 19th Century Pamphlets Online digitisation project. They are based on a particular form of publication (pamphlets) and particular equipment and requirements, but may be of use to others undertaking similar projects.

1. Book Scanners

1.1. Due to the fragile nature of the material, all pages were scanned by hand using the following book scanners within the Hartley Library scanning laboratory :

- PS7000 book scanners from Kodak[1]
- SupraScan book scanner from I2S[2]
- CopiBook book scanners from I2S[3]. These scanners are lit by two desk lamps with cold output
- Robotic Scanner from 4Digital Book operating in manual mode[4]

2. Image capture

2.1. Binding line margin

- Maintain at least 3-5mm between the binding line and the printed text. If this is not achievable on the CopiBook book scanners, pass the pamphlet on to the SupraScan scanner, which can scan deeper into the gutter. If the gutter is still not achievable, note this in the Scanning Database in the 'Binding' field.

2.2. Blank pages

- Scan all pages, including blanks. Scan the backs of fold-out maps to the same page size as their fronts.

2.3. Board, spine, front and back matter

- For pamphlets bound in original bindings, scan boards and spine in colour . Capture all front and back pages, which might include a handwritten table of contents. These images are saved within the following directories:
    - \master\library identifier\volume\front
      (for all front matter including the spine, front board and front matter)
    - \master\ library identifier \volume\back
      (for all back matter ending with the outer back board)

2.4. Damage

- Any physical damage to pages, board and spines should have been noted in the 'Notes' field within the Database by contributing libraries before the pamphlets were transferred. The damage should then have been verified by scanning laboratory upon arrival by the addition of the phrase 'Noted' within the 'Notes' field.
- Any damage not noted by the contributing library and any subsequent physical damage that occurs to a pamphlet, book or binding while at scanning laboratory, must be detailed in the 'Notes' field starting with the text: 'SCANNER NOTE ...'.
- Notes should be made of any tears that are >10mm and any damage that could lead to further damage. Reference should be made to the actual page number of the pamphlet and not the tiff number. The note should reference both the front and back of the page, for example 'page 4/5 tear noted'. Any damage caused during the scanning must be

[1] http://www.konicaminolta.co.uk/business-solutions/products/monochrome-systems/product-overview-discontinued-products/book-scanner-ps7000.html

[2] http://ww.i2s-bookscanner.com/en/products_SUPRASCAN.asp

[3] http://www.iiri.com/i2s/copibook.htm

[4] http://www.4digitalbooks.com/default.htm

noted, for example, 'SCANNER NOTE: Front board became detached during scanning process.'

2.5. Duplicates

- The libraries' preparations included checking for duplicates within their own collections and in other library collections. Only scan duplicates where a previous copy has been sent and marked as damaged.

2.6. File formats

- Files are saved in baseline TIFF 6.0 format.

2.7. Foldouts

- Scan foldout maps, diagrams or tables as one image. For 400dpi resolution can be maintained up to A1 (81.4" x 59.4cm). For paper sizes larger than A1, the dpi will be reduced to 300dpi.

2.8. Frame size

- Set a frame size for each pamphlet approximately 10-20mm beyond the three out edges and 10mm over the binding edge. The left and right frames should be identical in size and kept constant throughout the entire scanning of the pamphlet. For pamphlets with a large number of pages, the gutter may increase due to a relaxation of the binding. In these circumstances, the frame size should be large enough to accommodate this increase.

2.9. Glass use

- Pamphlet pages scanned on the PS7000 up to A3 size are scanned beneath 4mm float glass. This enables all page edges to be captured without any interference from fingers and to reduce rippling. With the PS7000 scanners, the glass is hand held horizontally across the pamphlet. The glass on the SupraScan book scanner is hinged and operates only in a horizontal position. Pages larger than A3 are scanned if appropriate with glass. The CopiBook scanners have a non-reflective glass plate which lowers automatically onto the pages.

2.10. Page masking

- Insert a thin black card (approximately 30cm square) beneath pages to provide a black outer edge. Insert the card up to a maximum of ten pages.

2.11. Rotation

- Scan all text in the reading orientation.

2.12. Scanning resolution

- All pamphlets are scanned at 100% (uninterpolated) using one of the following book scanners:
    - PS7000 Grey scale – 8 bit 400 dpi
    - SupraScan Grey scale – 8 bit or Colour-24bit 300 dpi
    - CopiBook Grey scale – 8 bit or Colour-24bit 300 dpi
    - DL Scanner Grey scale – 8 bit 300 dpi

- Pages containing only black and white print or writing should be scanned in greyscale. If a page contains any colour element (annotation, print, coloured paper) the whole page should be scanned in 24-bit colour. If the paper and print/writing cannot be distinguished in grey scale, the page should be scanned in 24-bit colour. Do not treat colouration due to aging as a colour element.

2.13. Scanning sequence

- Scan all pages in the order presented within the pamphlet or volume. If pages are bound out of sequence no attempt will be made to make a correction. If there appears to be missed page, a blank will not be added. If pages of a pamphlet have been mis-bound amongst or between other pamphlets, these pages will be scanned in the correct pamphlet page sequence, rather than the book order sequence.

2.14. Serial scanning

- Sometime are series of pamphlets have been treated by libraries as a serial publication and assigned a single catalogue record. In this case, all pamphlets in the series should be scanned sequentially.

2.15. Show-through

- Show-through can occur from four sources:
  - Bleed-through of ink from the back of the page
  - Over printing caused by the transfer of wet ink from the page opposite
  - Show-through of print from the back of the page
  - Show-through of print from the pages beneath due to paper transparency
- Show-through can be reduced at the scanning stage by the following processes:
  - On the PS7000 book scanners, reduce bleed-through by making a contrast adjustment.
  - Where there has been overprinting through the transfer of wet ink from the opposite page, use the SupraScan book scanner which has a 'softer' light source.
  - Where show-through is due to paper transparency, insert white paper beneath individual pages to enhance the contrast of the print.
  - Black paper beneath individual pages can be used to block show through from underlying pages.

2.16. Tables

- If a table or diagram goes across two pages, scan these as two separate pages.

2.17. Text alignment

- Pages are scanned to obtain relatively horizontal text. If pages are bound or printed on the skew, the scanner operator will shift the page to align the text horizontally.

2.18. Vertical binding alignment

- Where individual pamphlet pages are bound within a volume at differing vertical locations, the frame size should encompass the full extent of the vertical shift.

3. Metadata recorded during capture

3.1. Introduction

An in-house scanning database was utilised by the scanner operators to track the scanning of pamphlets, carry out QA and to record metadata items noted during the scanning process. Selected items are carried into the METS record.

3.2. Additional material

- If a pamphlet is accompanied by additional material, for example a letter, the tiff filenames of these pages should be noted in the Scanning Database within the 'Additional' field. The binding sequence should be maintained: for example, if the letter precedes a pamphlet, it is given the first filenames in the sequence. This metadata is included within the xml.

3.3. Annotations

- Annotations are defined as any additional word, letter, stamp, seal, underlining, embossment, intentional mark or additional print that has been added to a page subsequent to printing or binding. This does not include material added during a repair. If an annotation is present, the corresponding tiff number should be noted in the Scanning Database in the 'Annotations' field. The database makes no distinction between different kinds of annotation. This metadata is included within the xml.

3.4. Associated matter

- Associated matter is defined as material printed to accompany the pamphlet but which lies outside the main body of the text. Examples include adverts, petitions or invitations. Adverts within the body of the text should not be regarded as associated matter. The tiff filenames of all images containing associated matter should be noted in the Scanning Database within the 'Associated matter' field. This metadata is included within the xml.

3.5. Binding/Cropping

- For pages where a gutter of 3-5mm cannot be established, the individual pages should be noted in the Scanning Database in the 'Binding' field. Loose pages that are individually scanned and consequently have no gutter, should not be marked as a 'Binding/Cropping' issue. This metadata is included within the xml.

3.6. Clipping

- If any page clipping has occurred during binding that results in the loss of printed or hand written text, the corresponding tiff number should be noted in the Scanning Database in the 'Clipped' field. Loss of text due to clipping may have taken place along the three outer edges or bound into the gutter during binding. This metadata is included within the xml.

3.7. Colour

- The tiff filenames of all pages containing any colour component should be noted in the Scanning Database within the 'Colour' field. This will include coloured paper, text, images, maps, diagrams and annotations. This metadata is included within the xml.

3.8. Errata

- The tiff filenames of all pages containing errata slips or printed errata should be noted in the Scanning Database within the 'Errata' field. This metadata is included within the xml. Location of Addendum material should not be recorded.

3.9. Folded pages

- The tiff filenames of all pages containing any folded pages should be noted in the Scanning Database within the 'Folded' field. This metadata is included within the xml.

3.10. Foreign language

- For all complete pages within a pamphlet with text in a language other than English, the tiff filenames should be noted in the Scanning Database within the 'Foreign language' field. A drop-down box will enable a foreign language to be defined. English is the assumed default and no entry need be made to the database. This metadata is included within the xml.

3.11. Imagery

- The tiff filenames of all pages containing any printed imagery associated with the pamphlet text and provided by the author should be noted in the Scanning Database within the 'Imagery' field. This should not include printer's embellished letters or lines, or arrows. This metadata is included within the xml.

3.12. OCR issues

- Any bleed through, show through, wet ink transfer, marks, or foxing that is likely to limit the OCR of the printed text should be noted in the Scanning Database. List the tiff filenames within the 'OCR issue' field. This metadata is included within the xml.

3.13. Pagination

- The tiff filenames for all pages within a pamphlet that are incorrectly paginated (due to either printing or binding errors) should be noted in the Scanning Database within the 'Pagination' field. This metadata is included within the xml. Pages should be scanned in the order of the pamphlet, and no corrections made. If an incorrect pagination affects the remaining pages of a pamphlet, all the pages following the error should be noted in the database. Where a pamphlet uses a numbering sequence not starting at zero (for example beginning on page 267) and the sequence is correctly maintained beyond this, this should not be marked in the 'Pagination' field.

3.14. Rotation

- Pages printed or bound in a landscape position, for example tables and images, are rotated by the scanner software at the time of scanning to a read orientation. Images are then saved in this read orientation. The tiff filenames of all pages rotated should be noted in the Scanning Database within the 'Rotation' field. This metadata is included within the xml

3.15. Tables

- The tiff filenames of all pages containing any tabular row and column data, or list data, or bulleted lists should be noted in the Scanning Database within the 'Table' field. This metadata are included within the xml. Do not include Table of Contents or indexes or hand-drawn tabular data, but include worked equations. This metadata is included within the xml.

3.16. Watermarks

- These should be noted in the 'Notes' field using the standard text 'water marks'

4. Post-processing

- Prior to OCR the images are automatically rotated with proprietary software to give a horizontal text-block and resaved in this form. Due to the nature of the printed text, this does not mean that every line will be completely aligned or straight.
- Images are then cropped by the same proprietary software to the three outer edges of the page and within the binding line. This will result in some black edging included with the images.
- Images are then saved with LZW compression.