Searching scanned documents

Extracted from advice in biographic research for bellringers – John Harrison, Feb 2019

Increasingly, historic documents are being scanned to produce an image that can be viewed digitally, and in many cases (but not all) the image has also been analysed using OCR (Optical Character Recognition) to generate a digital text file that can be searched electronically for words or phrases. You don't normally see the converted text because it is hidden behind the scanned image of the page. But when you appear to be selecting and copying what you can see, you are actually selecting and copying the corresponding text that is hidden behind the image.

When you search the document digitally you get a match if (and only if) the text you are seeking is in the hidden converted text. If the OCR is perfect, then the hidden text is the same as the image that you see. For the vast majority of scanned words it is – but OCR isn't perfect and occasional characters can be misinterpreted. If a misinterpreted character is in a word you are looking for then the search won't find it because it won't match.

Some real examples are:
ful! (full), o f (of), cau.se (cause), N e w o a s tle (Newcastle), WoonPToCK RO'D (Woodstock Road),
Sp.ce (Spice), MINEHIAD, 80MERSET (MINEHEAD SOMERSET), practical!y (practically).

Words or phrases that span a line break appear as separate parts that might not be recognised together, especially in a multi-column document if the OCR text isn't structured in corresponding blocks.

The OCR result can be proof read and corrected, but that is expensive so not usually done. OCR software has improved over the years – more recently scanned documents should have far fewer (but not zero) errors.

OCR errors undermine the reliability of digital searches. Finding something means it is there, but not finding it doesn't mean it's not there. It could be there but with one or more corrupted characters.

To improve the chances of a hit you can try using a partial search term, and then visually scan** the results to eliminate any that aren't relevant. For example if you are looking for 'Williamson' you could try searching for 'Willia' or 'iamso'. The first will also find 'William', 'Williams' and McWilliam but the latter will probably only find 'Williamson'.

A (real) example of using this technique was searching for Minehead in one pre-war year of The Ringing World. The search term 'minehead' gave 28 hits but the search term 'mineh' gave 47 hits (including several 'MINEHIAD', 'MINEHIAO', 'MINEHtAO' and 'MINEHiA').

Modern documents that have been electronically produced don't rely on OCR because the text is normally embedded in the original document. That makes searching more reliable, so (give or take any spelling mistakes) you should be able to find every instance of what you are looking for.

A later practical example:

So far I have found the following versions of 'kirkby' :

k ir k b y

k irkby

k jrkby

kirby

k irby

k irkl b y

k irk ejy

KIRK BY-IN-ASIIFIFXD

KIRK BY -IN-ASH FIELD

K 1R K B Y -IN -A S H F IE L D

K IR K B Y -IN -A S H F IE L D

K IR X B Y -IN -A S H F IE L D

K IR K B Y -IN -A S H FIE L D

K IR K BY-IN- A SH F IE L D

K IR K Y -IN -A S H F IE L D

K IR K B Y -IN -A S H FIE L D

K 1R K B Y -IN -A S H FIE L D

K I R K B Y -I N - A S H F I E L D

No wonder my initial naive search for 'kirkby' missed a lot of things I knew should be there.