Avoiding problems in OCR

As the foregoing will have made clear, one of the chief contributing factors to the success of the OCR process is the quality of the initial scan. Three broad factors external to the quality of the OCR software itself can have potential negative effects on high OCR accuracy. These are:

- The underlying qualities of the source material
- The initial image capturing process (including the technology used to create the image)
- Image enhancing techniques employed to improve legibility/readability

This section deals with these common barriers to high OCR accuracy and how to avoid them. Note that negative factors resulting from source material may not be correctable, though they may guide selection of material for OCR.

Negative factors resulting from the source material

- **Yellowed paper**
  
  Background colour may change across document due to age of material, causing fuzziness at edges of characters. This may have a minor effect on OCR accuracy.

- **Warped paper**
  
  The combination of humidity and age means that the pages of many books are warped rather than flat. As a result, the text warps too. This will have a relatively high impact on OCR accuracy, especially if found in combination with bad printing.
- **Bleed through**

  Bleed through occurs when two leaves of a book have been pressed together before the ink dries. This can have very negative effects on OCR accuracy.

  Image taken from the collection of the Bayerische Staatsbibliothek, as part of the IMPACT random dataset. Reprinted by permission.

- **Show through of ink**

  In documents printed on very thin paper, the ink from one side of the page will often show through to the other. This can have severe negative effects, comparable in all ways to bleed through.

  Image taken from the collection of the Bayerische Staatsbibliothek, as part of the IMPACT random dataset. Reprinted by permission.

- **Bad printing**

  Blurred, fat, broken or faded characters. These can have highly negative effects, particularly where characters are blurred together or broken into more than one shape.

  Blurred
Broken and dotted

![Image](image-url)

Fat

Images taken from the collection of the Bayerische Staatsbibliothek, as part of the IMPACT random dataset. Reprinted by permission.

- More than one ink colour
  Moderately severe effects can occur. OCR may not recognise complete characters or words.

![Image](image-url)

Annotations

These can include notes and drawings by users; also library stamps and watermarks. These are very unlikely to be read correctly, but a deeper problem is that annotations can confuse the segmentation process of OCR engines (e.g. the means by which it identifies character/word/line blocks).

![Image](image-url)
- Lack of lexicon data

This occurs when the OCR engine does not have access to relevant language data for the document. This leads to unrecognised words, and perhaps more damagingly to false substitutions of words.

Negative factors resulting from the image capture process

In general, the best OCR results will be produced by creating images at a resolution of 400 pixels per inch (ppi) or above, in colour or grey-scale.¹ These standards will preserve the vast majority of the detail in the original item, where lower resolutions will result in progressively worse OCR recognition.

However, capturing images in such high quality has cost and storage implications, particularly if the master image file is to be archived indefinitely. Some institutions engaged in mass digitisation create an OCR output from the master file, and replacing the master image with a compressed access copy to conserve storage space.

These are the most common ways in which image capture can negatively effect OCR accuracy:

- Narrow binding

May result in geometrical distortion. This may result from carelessness by the camera/scanner operator in correcting for the tightness of the binding, or from the inability of a particular scanner to deal with books that will not comfortably open beyond 60°. It can have very high negative consequences on OCR accuracy.

- Image not cropped

Borders and unnecessary detail from facing page may appear in scan. It is particularly prevalent in scans from microfilm. While very common, it will generally have a low effect on OCR accuracy; most OCR engines can correct for these details.

¹ The superiority of greyscale over bitonal scanning has been called into question in a recent study:


Image skew

In mass digitisation conditions, the scanner will not be adjusted page-by-page for the relevant print space (i.e. just that part of the material that is needed). This can lead to the creation of skewed images. Because a common problem, OCR engines have been trained to correct for it, so effects on OCR accuracy are generally low. However, the severity of the effect on OCR depends naturally on the degree of skew.

Image taken from the collection of the Bayerische Staatsbibliothek, as part of the IMPACT random dataset. Reprinted by permission.
Factors resulting from image enhancement techniques

Most industrial image capturing technology is bundled with a suite of image processing and enhancement software. Some of the above-mentioned factors that negatively effect OCR results can be “corrected” by this software, so that a skewed image can be reoriented, or a page automatically cropped so no unnecessary detail is included in the image file, etc. However, the results of these image enhancing tools are often difficult to objectively judge, because the software works only with its parent scanner. The effectiveness of the tools is therefore directly related to the initial effectiveness of the scan.

In addition to this, image enhancement tools of this type often feature a range of advance features that allow images to be altered post-capture, so page images can be automatically sharpened or have their contrast boosted, etc. But problems can occur when images are enhanced in a large batch: while the OCR accuracy from some images will be enhanced by the application of image sharpening, some could be negatively affected.

OCR engines also include as standard a number of image enhancement tools. The most common are:

- **Binarisation tools** - where a colour or grey-scale picture is converted into a pure black and white (binary) image. Making an image bi-tonal in this fashion helps the OCR engine to separate textual detail from background detail. It will also reduce the effects of bleed-through and show-through;
- **Straightening** - which corrects the angle of a page;
- **Dewarping** - which corrects for the curve of a page and the characters on it;
- **Border removal** - which removes the borders at the edge of a scan. This minimises the storage space required by a particular image and can also improve OCR – seeking as it does to ensure that no extraneous details are recognised as characters;
- **Page segmentation** - which can divide a page image into headers and text, individual paragraphs and columns and, in the case of newspapers, into individual articles.

The image enhancement software in scanners and OCR engines can produce unwanted effects when not set up with enough care:

- **Bitonal output reduces readability**

  Scanners and OCR engines will automatically transform grey or colour images into binary images before OCR. If a document is particularly difficult or if the parameters for binarisation are not set with enough care, OCR accuracy can be severely reduced;

Image taken from the collection of the Bayerische Staatsbibliothek, as part of the IMPACT random dataset. Reprinted by permission.
• Under-correction of processing software for poor quality scan

The image processing software in a scanner or OCR engine may not be sufficiently sophisticated to correct for an image that has been captured with insufficient regard to lighting, contrast, etc. This is particularly true of bitonal image creation and often results in broken text. This can have very high negative effects on OCR.

Image taken from the collection of the Bayerische Staatsbibliothek, as part of the IMPACT random dataset. Reprinted by permission.
Niall Anderson joined the IMPACT Project as a full-time Project Analyst in October 2008. He has worked for the British Library since 2002 in a variety of roles, most relevantly as a digital content/metadata editor for the Collect Britain and Archival Sound Recordings digitisation projects. He has also worked for IBM as a communications officer and technical writer.

### Revisions

<table>
<thead>
<tr>
<th>Version</th>
<th>Status</th>
<th>Date Released</th>
<th>Lead Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>Pilot Release</td>
<td>14.10.2010</td>
<td>Niall Anderson – British Library</td>
</tr>
</tbody>
</table>

### About this Document

This document is the second section of the IMPACT: Best Practice Guide to OCR. There are two more sections within this Best Practice Guide, which supports the IMPACT: Briefing Paper on OCR.

This Best Practice Guide has been released by the IMPACT project to assist practitioners and students in the mass digitisation of text and the use of Optical Character Recognition.

It is currently in a draft form and has been pre-released for public comment through our LinkedIn group, with a closing date for feedback on the 26th of November 2010.

You can help us to improve these materials by leaving your comments in the discussion area of the IMPACT Improving Access to Text LinkedIn group which is now a public group. If you would prefer to comment in private, please email your comments directly to feedback.impact@gmail.com

IMPACT will be gradually releasing a wide range of materials to support mass digitisation and OCR – which will all be shortly available through the IMPACT website at: [http://www.impact-project.eu/](http://www.impact-project.eu/)