



Koala White Paper:  
**OCR and Quality Assurance**

KOALA

Koala Publishing Ltd.  
Downend House, 112 North Street  
Downend, Bristol BS16 5SE  
United Kingdom

Tel: +44(0)117 910 9111  
[www.koalapub.co.uk](http://www.koalapub.co.uk)

## Contents

<b>Background.....</b>	<b>3</b>
<b>Factors Affecting the Level of Accuracy.....</b>	<b>4</b>
<b>Putting a Value the Level of Accuracy.....</b>	<b>5</b>
<b>Improving the level of Accuracy.....</b>	<b>6</b>
<b>Ability to Replicate the Results.....</b>	<b>7</b>
<b>Logging the Actions of the User for subsequent Audit.....</b>	<b>8</b>
<b>And finally .....</b>	<b>9</b>

## Background

We have recently been conducting testing and research into OCR (Optical Character Recognition), assessing the strengths and weaknesses of various systems.

Our objective was to find the 'best-fit' system and/or procedure to enable one of our customers to convert a variety of disparate paper copies and bitmap images of technical documents into a regulated, structured xml format. The nature of the data meant that accurate, high quality recognition was a prerequisite and the ability to audit the whole process was high on the wish list as well. Simple? Well so we may have thought!

Now OCR can be very accurate indeed, especially when reading pure text documents from good clean copy or scans. And if the layouts on each page are simple and consistent, that helps even more. However, our customer's source documents for OCR were going to be:

- Produced by different organisations - so in various layouts
- Often copies of copies, including old fax machine prints and bits of paper physically pasted together
- Combinations of page sizes and orientations
- Pages with complex and inconsistent table structures
- Technical data with code sequences, numbers, map references, radio frequencies, abbreviations, symbols and so forth
- Containing text in colour, often on coloured or tinted backgrounds.

Despite all of the above challenges, we found that modern OCR systems handled, on average, a good 85% of this difficult content very well indeed and, for the requirements of this specific customer, one system (Abbyy FineReader) achieved around 95% accuracy when the appropriate options were selected for each document.

Now the question arose: "How does the process sit with QA requirements?"

## Factors Affecting the Level of Accuracy

Evaluating how accurate and comprehensive an OCR operation has been is challenging.

Many systems, including FineReader, will highlight parts of the content which they are unable to 'read' or about which they are uncertain. Some may even enumerate these for you. However it is inevitably up to the user to manually go through and type in what they 'believe' to be the correct character or character string for each instance. Note the use of the word 'believe' in the previous sentence - with blurred or feint originals, it's often difficult for even an experienced expert in the subject matter, to be absolutely certain of a word or number. We also have to bear in mind that many of the characters highlighted as 'uncertain' will in fact be correct, whilst others which are not highlighted may be in error.

We can employ spell checking but, as we all know, that only covers words saved in the software's dictionaries, not necessarily your technical terms and certainly not codes and numeric strings. And if an 'm' is read as 'rn' (a common problem with blurred originals, low resolution scans and small font sizes) a 'chum' becomes a 'churn' and no spell or grammar checker is going to pick that up!

So, the ONLY answer, if you need a high level of accuracy, is to get the whole of the text proof read, character by character, by an expert in that subject, who is experienced in professional proof reading. And who is also an expert in the language the document was written in!

## Putting a Value the Level of Accuracy

How can we put a quality value on the results of an OCR conversion?

- By the number of characters flagged by the software?
- The number of strings of characters manually corrected?
- The number of words corrected?
- The number of characters corrected?
- The amount of time it took to proof read and correct the document?
- Do we put a different 'weighting' on the accuracy of some parts of the document (for example, those containing safety-critical data) compared to others?

It all depends. We probably need to combine two or more of the above criteria and formulate a scoring system according to our requirements.

As most of the accuracy checking is inevitably going to be a manual process, it follows that the number of errors actually spotted will not necessarily be 100% and that no amount of checking will ever guarantee to make it so.

Recording the number of errors spotted and/or corrected is also usually a manual process. We could however take a copy of the file after the initial OCR read process and compare that to the file after editing is complete and then run a compare program to count the number of characters deleted, edited and added.

## Improving the level of Accuracy

After a significant number of 'typical' documents have been processed, some average values can be established.

Whether these are acceptable or not will depend on the nature of the documents. For example, some internally archived administrative documents may be acceptable with just 80% accuracy, especially if the originals (and/or bitmap scans) are retained and can still be easily located when needed. However, for safety-critical, regulatory and most externally distributed documents, a much higher level of accuracy must be sought.

Experience will dictate the most reliable OCR methodology and settings for obtaining the optimum results with the minimum level of manual processing. OCR systems can be fine tuned and put through a 'learning' phase to help recognise new character shapes. This is typically used with unusual or elaborate font types. However, in our experience it's a lengthy process and, for the general run of documents, only improves the accuracy rate by less than 1%.

## Ability to Replicate the Results

Whilst it is a normal QA requirement to be able to demonstrate quality by replicating results on a second pass through the system, this procedure is not always satisfactory for OCR. Even using the exact same scan(s) as a starting point, and the same OCR software with the same configurations and options, subsequent 'reads' will often produce slightly different results. If you then add the 'human factor' of the editing and checking procedures, it has to be accepted that it's very unlikely that results are going to be exactly the same on two or more passes of a sizeable document

## Logging the Actions of the User for subsequent Audit

For most projects, this too has to be a manual process, the operator entering details of software used, files processed (and their locations), options selected for each document processed, times, user IDs and so forth. Software can be developed to record some of this information automatically and to archive the file sets created but this is rarely justifiable in cost terms.

In respect of auditing the actual edits made, as previously stated this can be done by capturing a 'snapshot' of the OCR output before and after editing. A more detailed auditing of the edits would not be realistic.

If money and storage collateral were not considerations, the ultimate audit trail would be to record every keyboard stroke and mouse click together with a video of the screen display for every document processed.

## And finally ...

This project has been a very thought provoking undertaking. It has reinforced the fact that OCR is:

- Always at its best when set up for consistently styled and structured quality scans
- Almost always less time consuming than re-keying
- Very powerful but not 100% reliable, regardless of how good the original source material is
- A process that cannot be left entirely to the software (unless you are happy to accept an arbitrary error rate)
- Always going to require reasonably high levels of user input to get accurate results from difficult content
- Very challenging to audit!
- Nevertheless, getting better year on year - worth investigating all new releases.