

## Optical Character Recognition (OCR) on Newspapers – An Overview

Author: Rose Holley (Manager – ANDP)

Version: 1.0

Date: 26 September 2007

### 1. Overview

This document outlines important aspects relating to OCR and aims to answer common questions about the process.

Optical Character Recognition (OCR) is a process run by OCR software. The software will open a digital image e.g. tiff file containing full text characters and then attempt to read and translate the characters into recognizable full text and save them as a full text file. This is a quick process that enables automated conversion of millions of images into full-text files that can then be searched by word or character. This is a very useful and cost efficient process for large scale digitisation projects involving books, journals and newspapers. There are several OCR software packages on the market but a popular package for older material or that in languages other than English is Abbyy Finereader. This is currently being used by several newspaper projects world wide.

The OCR process is dependant upon a number of factors and these factors influence results quite radically. Experience to date has shown that using OCR software over good quality clean images (e.g. a new PDF file) has excellent results and most characters will be recognized correctly therefore leading to successful word searching and retrieval. However over older materials e.g. books and newspapers the OCR is extremely variable and for this reason some projects advocate re-keying the text from scratch, rather than attempting OCR. Contractors who offer re-keying and OCR services are mostly based in India. The process is staff intensive and sometimes a combination of both re-keying and OCR will be performed for a project. It is usual to undertake sample tests on the actual source material to be digitised before making decisions about OCR and re-keying.

### 2. Factors Effecting OCR results

There are several things that affect the results as outlined below:

#### 2.1 The source

Within the original material

- Highly complex layout
- Radical differences in layout over time
- Variable font sizes and character types (especially Gothic)
- Narrow space between lines
- Narrow gutter between columns
- Missing or misprinted text
- Poor quality or deteriorated inks
- Poor quality or deteriorated papers
- Irregular alignment of characters in hand-set press
- Annotations by hand
- Graphic devices and/or elements
- Mixed languages/fonts on the same page

From microfilm

- Poorer quality in second or third generation copies if not using master
- When film made the lens was not focussed correctly so text is not sharp
- Skewed (i.e., curvature of text from gutters within bound volumes) or twisted lines of text. Skew at page and column level.
- Noise, dirt, scratches
- Text cut off completely due to poor filming/tight gutters
- Broken lines, broken characters
- The age of the microfilm (anything over 20 years old has very variable density throughout the reel which means auto adjust programs for the whole reel do not run effectively)

## 2.2 Supplied Source

Although a human can get a rough idea of whether the source they are supplying will be good for OCR, in most cases it is difficult to tell exactly. The 'OCR' eye looks at images in a very different way to the human eye and often what a human considers bad for OCR is excellent and what would thought to be good yields bad results. (The ANDP is having a large variety of source scanned in the pilot phase because of this difficulty of the human eye being unable to tell what will be good or bad).

## 2.3 Image manipulation

Good OCR depends largely of the level of image manipulation which has taken place on the image before the OCR process commences. Currently image manipulation normally takes place in either the scanning software or stand alone software programs before OCR. Most OCR programs offer only basic or no image manipulation before OCR commences. It is preferable to do the following types of image manipulation before OCR (these cannot be done effectively in Photoshop they do require a specific image manipulation pieced of software).

- De-skew page/column to within +/- 1
- Remove noise/dirt/scratches
- Increase contrast of black on white (bi-tonal files)
- Characters – smoothing, rounding, sharpening etc
- Adjusting density levels

## 2.4 Image type

It is preferable to use tiff images rather than jpg images since they contain more pixel information. A jpg is a compressed file that has 'lost' pixel information in the compression process.

There is no consensus to date from contractors or information professionals about whether bi-tonal or greyscale gets better OCR results. It depends what software is being used.

## 2.5 Dictionaries

When OCR is being performed it is possible to use dictionaries against which words are cross checked. Most OCR software has complicated algorithms which are utilised when the software is unsure of character recognition. Abbyy Finereader has the capacity for several user defined dictionaries to be loaded with for example peoples or place names, or different languages.

## 2.6 Training

Some OCR software systems (e.g. Abbyy) have built in 'intelligence' or have the capability to be 'trained'. What this means is with some manual intervention of OCR the system will improve its rate of accuracy as time goes on over the same source i.e. it learns how you want it to treat certain occurrences. This also speeds up the OCR process. This training was notably done on the Maori newspaper digitisation program using Abbyy Finereader.

## 2.7 Manual Intervention

With manual intervention (i.e. an operator makes the decision on what unrecognised characters should actually be) OCR accuracy can be perfect no matter how bad the source. Unfortunately this is an extremely time consuming process and in most cases unrealistic for large scale digitisation projects.

### 3. The ANDP Process

The ANDP has contracted a supplier in India to deliver OCR text. Apex CoVantage is currently performing OCR over all newspapers using Abby Finereader OCR software. In addition the headings of articles and first 4 lines are being manually checked and corrected by hand to ensure that these are 99% accurate. No re-keying is being undertaken. The uncorrected and corrected data for the article titles and first 4 lines are being returned to NLA so that further research can be done. No system 'training' is taking place. A NLA user defined dictionary is being applied containing geographic place names.

Files being provided to Apex for OCR are 'image optimised' bi-tonal tiff files. Newspapers have been scanned as both bi-tonal and greyscale tiff files at the scanning bureau. The scanning software (NextStar) then carries out auto manipulation to optimize files for OCR (e.g. smoothing characters, sharpening, removing dirt and noise, increasing contrast). NLA has chosen to receive both bi-tonal and greyscale files from the scanning bureau since it is still unclear in the future which will be the preferred file for best OCR. To date most OCR contractors are specifically recommending one or the other but this is not consistent amongst contractors.

### 4. OCR accuracy

#### 4.1 Measuring Accuracy

When files have been processed it is desirable to know what accuracy rates certain sources produce. This is because % accuracy directly affects success of search results. It is commonly agreed that 'good' accuracy needs to be above 98%. Accuracy can be measured in characters or words. Measuring a file using both these methods will usually give very different results. A file could be a whole page or just an article. Unlike books and journals newspaper pages cannot be compared to each other because of the varying layout on pages so it is hard to give an overall accuracy figure for a digital collection. In addition it is hard to give an accuracy level for an individual newspaper title since layout changes over time, and varying qualities of source have been used. ANDP are still doing further research on OCR accuracy levels and have not reached any significant results for Australian newspapers, largely due to the variation of factors outlined above – all of which have been relevant in the pilot sample of 50,000 files.

OCR accuracy can be measured by hand i.e. 'proofreading' or by automated programs. To date NLA and Apex CoVantage have been checking accuracy by hand. This is a time consuming process and takes on average 8 hours for a complete newspaper page of approximately 40K in size.

#### 4.2 ANDP Accuracy results

ANDP chose a representative sample of 30 pages split into 3 groups from the first 50,000 pages to be processed. The files were bi-tonal and the Abby default dictionary was used. The results were measured on character accuracy and are:

Average character accuracy:

Good group	98.02%
Average group	92.61%
Bad group	71.00%

Some individual bad pages scored only 5% accuracy. The ANDP do not consider these results to be conclusive or necessarily accurate in view of later information received, and will be doing further research in this area with a view to improving OCR accuracy by a number of means.

### 5. Revealing OCR text to users

It has yet to be decided if there is any benefit in revealing the raw OCR text to users. This largely depends on the levels of accuracy obtained. When files have a low accuracy rate the text will look like 'gibberish'. Fields that will be appearing in search results are the article headings and possibly the first 4 lines of text. Since these fields have been 'cleaned' to 99% accuracy these fields are helpful to the user.