



Designing documents for automatic recognition

Guidelines scanning

Service Point Belgium

Author	Johan Rambags
Date	30-06-2009
Version	2.0



Table of Contents

1	Introduction.....	2
2	General guidelines	2
2.1	Font type.....	2
2.2	Color dropout.....	2
2.3	Quality of the scans	3
3	Automatic data recognition.....	3
3.1	The importance of a well designed document layout	3
3.2	Elements on the form	3
3.2.1	Form identification.....	3
3.2.2	Anchor points.....	4
3.3	Data on the document	4
3.3.1	OCR	4
3.3.2	ICR.....	4
3.3.3	OMR.....	5
3.3.4	Barcode.....	5
4	Annex: drop out colors	6
4.1	Drop out color red.....	6
4.1.1	Pantone yellow.....	6
4.1.2	Pantone orange	6
4.1.3	Pantone red/pink/peach.....	6
4.1.4	Pantone fuchsia/purple	6
4.2	Drop out color green.....	7
4.2.1	Pantone yellow.....	7
4.2.2	Pantone orange	7
4.2.3	Pantone green/blue.....	7
4.2.4	Pantone yellow/green	7
4.3	Drop out blue	7
4.3.1	Pantone blue.....	7



Service Point

Document, Print and
Information Management

1 Introduction

The results of automatic recognition of documents largely depend on the design of those documents. Issues such as font type, usage of colors and layout can greatly improve the results of that automatic recognition of document data when correctly applied.

This document provides an overview of guidelines and best practices to improve results of automatic recognition by designing scan friendly documents.

2 General guidelines

2.1 Font type

To achieve good results with automatic recognition, one of the key factors is to make sure that the characters are easy to read. That means for recognition software that there is enough white space around the characters. Too little white space around characters leads to recognition errors. It is therefore recommended to use one of the following font types:

- arial
- century gothic
- courier new
- lucida console
- times new roman
- verdana

If another font type is used, make sure that enough white space is around the characters. Do certainly not use a "fantasy font type". Try to limit the number of different font types, if possible only use one font type.

The second characteristic of a font that is important is the font size. To get good recognition results, the minimal height used is 10 pt. It is recommended to use the same height as much as possible and certainly within one data zone.

Some other recommendations:

- print characters in upper case
- use one color for the data that needs to be processed, preferably black
- don't print characters in italic, underlined and bold

2.2 Color dropout

Current (high volume) production scanners support color dropout. The most used dropout color is red, but also green and blue are used. An overview of the colors that can be used for dropout can be found in the annex.

Borders, zones and other lines disturb automatic recognition, because the software does not know the difference between actual data and preprinted lines. By using dropout colors to print borders and zones on a document the person who is filling in that document sees those lines and can fill in that document correctly. During scanning however, the dropout color is filtered from the image (is not visible on the scanned image) and the borders can no longer interfere with the recognition.

Dropout colors are widely used, for example on tax forms and money transfer forms. Below is an example of an OMR field that is printed in a green dropout color:

Geslacht

M V

It is important however to keep in mind that when a dropout color is used, that all data and information in that color is filtered out. That means that when for example a red color is used for dropout, all text written in red ink will also be filtered. It is recommended to specify on the document which colors of ink a person can use to fill in the document.

2.3 Quality of the scans

As the scanned image is the start of the flow for recognition, the quality of those images is very important. The minimum resolution needs to be respected and it is recommended that quality enhancement (deskew, black border removal, rotation etc) is used.

For automatic recognition, use following minimal resolution:

- 300 DPI for black & white and grey scale images
- 150 DPI for color images

3 Automatic data recognition

3.1 The importance of a well designed document layout

The software that performs the recognition needs to be configured to extract the data. To achieve good results with recognition, several elements need to be present on a document. Based on those elements the recognition software knows what the type of document is and where the zones are located that contain the data. Placing those elements correctly on the document is key to getting good recognition results.

A good document layout is not only of importance to recognition software. At least as important is that the person filling in the document can do this in an easy way and that he/she is guided as much as possible when entering data. For example a date of birth: two character boxes need to be created to enter the day of birth, two boxes for the month and four for the year. In this way the user knows the requested date format and on top of that the user will not try to enter a slash or a hyphen as a separator between the different parts of the date.

3.2 Elements on the form

3.2.1 Form identification

By placing a unique identification on a document, the recognition software can more easily determine what form (and as a consequence what data) needs to be processed. When multiple document types need to be processed at the same time (in the same batches), document identification is mandatory.

Several solutions are available to create identification on a document. The best option is to use a barcode, but pre-printed text is also a possibility. The most important condition is that the value of the identification is unique within all types of documents that processed together at the same time. A second condition is that the location where the identification is placed on the document is always the same for all documents of that type. More guidelines on using barcodes and pre-printed text can be found under paragraph "Data on the document".

3.2.2 Anchor points

When the recognition software has determined what type of document needs to be processed, it looks for the anchor points. Anchor points are fixed characters or symbols on a fixed location on the document and are used by the recognition software to calculate where the actual data zones are on the document.

Anchor points can be printed as a symbol or as a text and preferably a combination of the two. Each document needs at least three anchor points which should not be on one line.

3.2.2.1 Symbols or text

- Conditions for using a text as anchor point:
- comply to the general guidelines for font types
- characters should be chosen from the following list: A B H J K M N O P R S T X Y Z or use numerical characters
- minimum 3 mm white space around the text.

3.2.2.2 Crossing lines

Lines used as anchor points need to comply with several conditions:

- minimum 10 mm length.
- approximately 1 mm line thickness.
- minimum 10 mm from the edge of the document

Example of two anchor points (left- and right upper corner):



It's not always necessary to create separate lines or crossings on a document. Already existing lines that comply to the conditions above can also be used.

3.3 Data on the document

3.3.1 OCR

To get good results from recognition of pre-printed text, the following rules should be applied:

- comply to the general guidelines for font types
- leave at least 3 mm of white space around the data zone
- use as much upper case characters and numerical characters and avoid symbols as the slash, hyphen etc when possible

3.3.2 ICR

Recognition of handwritten texts is still under development and although results are being improved continuously, they are not yet as good as for OCR. To get the best results, following can be applied:

- use a separate box for every character that needs to be entered



4 Annex: drop out colors

The colors mentioned below are recommended in combination with red, green and blue color filtering. Although these colors provide good results in normal circumstances, it is recommended to perform tests on the scanner that will be used during production before sending the documents to the printer.

4.1 Drop out color red

4.1.1 Pantone yellow

- 100	- 106	- 109	- 115
- 101	- 107	- 113	- 116
- 102	- 108	- 114	- Pantone Yellow *

4.1.2 Pantone orange

- 120	- 123 *	- 129	- 135
- 121	- 127	- 130 *	- 136
- 122	- 128	- 134	- 137 *

4.1.3 Pantone red/pink/peach

- 1345	- 141	- 148	- 155
- 1355	- 142	- 149	- 156
- 1365 *	- 143 *	- 150 *	- 157
- 1555	- 162	- 1625	- 169
- 1565	- 163	- 1635	- 170
- 1575 *	- 164 *	- 1645 *	- 171 *
- 176	- 1765	- 182	- 189
- 177	- 1775	- 183	- 190
- 178 *	- 1785 *	- 184 *	- 191 *
- 196	- 204	- 217	- 223
- 197	- 210	- 218	- 224
- 198 *	- 211 *	- 219 *	- 225
- 487 *	- 488	- 489	

4.1.4 Pantone fuchsia/purple

- 230	- 236	- 243	- 250
- 231	- 237	- 244	- 251



Service Point

Document, Print and
Information Management

- 232 * - 238 * - 245 * - 252 *

* acceptable results when used as "tint"

4.2 Drop out color green

4.2.1 Pantone yellow

- 100 - 106 - 109 * - 114
- 101 - 107 - 113 - Pantone yellow *
- 102 - 108

4.2.2 Pantone orange

- 120 - 127 - 134 - 1345
- 121 * - 128 *

4.2.3 Pantone green/blue

- 372 - 379 - 386 - 389 *
- 373 * - 380 * - 387 - 565 *
- 374 * - 381 * - 388 - 566 *

- 572 * - 586 * - 580 * - 5807 *
- 573 * - 5865 * - 587 - 331 *
- 585 * - 5875 * - 5803 * - 332 *

4.2.4 Pantone yellow/green

- 393 - 395 - 3935 - 3955
- 394 - 396 - 3945 - 3965

* acceptable results when used as "tint"

4.3 Drop out blue

4.3.1 Pantone blue

- 250 - 283 - 2975 - 310
- 263 - 290 - 304 - 3105
- 2635 - 297 - 305 - 317
- 277