# ScanMan3 Document Filing Software

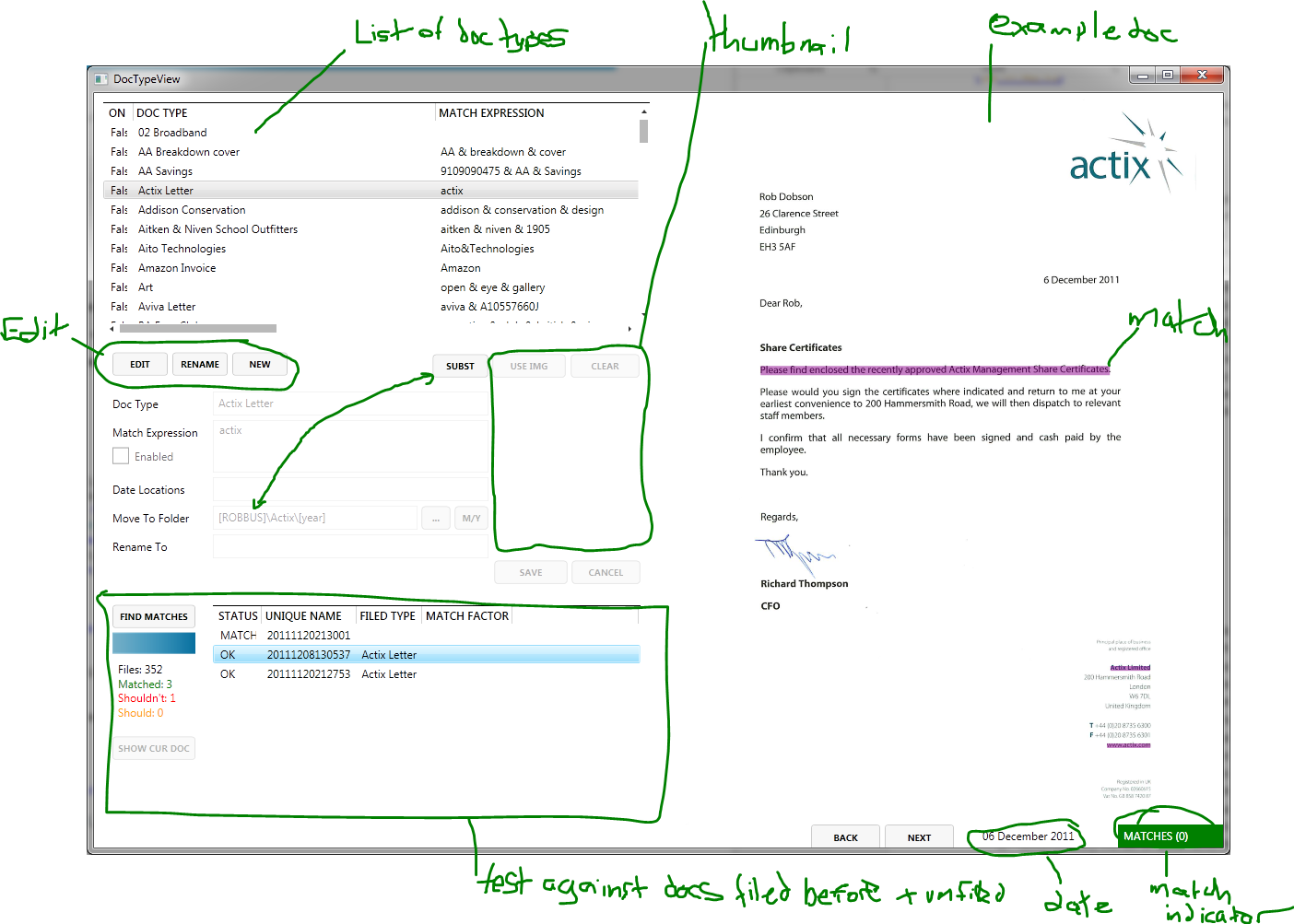
Version 0.1 11th March 2014

# Overview

This software is designed to help recognising and filing scanned documents.

Currently the main task is to “teach” the system how to recognise different kinds of documents reliably. To do this the part of the software relating to creating, refining and testing Document Types has been substantially completed. It now needs thorough testing and usage in order to create the set of Document Types that we will use in future for our filing.

The part of the software used for managing Document Types currently looks like this:



# Getting Started

To get started look at the DocTypeView (currently this is a button on the opening window). The list at the top right is all currently defined Document Types – these are imported from the previous software and many / most are not well defined at all. They are initially disabled (the Enabled check box is clear) so that they will not be used when matching against documents. In order to enable a Document Type it is necessary to press the Edit button (this is referred to as going into Edit Mode in the rest of this document). The Match Expression and other fields can now be changed.

To test any changes press the FIND MATCHES button. The Document Type will be compared against all previously filed documents and unfiled ones. The potential results of this operation are:

* OK – the document is matched to the same Document Type as when it was filed
* MATCHES-BUT-SHOULDN’T – the document matches the current Document Type but was filed under a different type previously. This may indicate that the current Match Expression needs more work but it could also indicate that the document was previously incorrectly filed or that the document contains pages from different kinds of document (this has happened in the past due to careless scanning) OR it might indicate that the Document Type name has been changed.
* SHOULD-BUT-DOESN’T – the document in question should match (it was previously filed with the current Document Type) but it doesn’t match. Many of the gotchas described above can apply in this case too so it needs careful consideration.
* UNFILED – the document is unfiled so the software has no idea what Document Type it should match to.

It’s important to work the match information to make sure that the new Document Type is as well defined as it can reasonably be.

## Example File Display

Once one or more documents have been found in the FIND MATCHES testing process they can be clicked on in the results list and the document in question will be shown. At this point the following information will be shown on the right side of the window (on and below the document page):

* Location rectangles that may be used in the Match Expression and Date Locations boxes
* Matched text boxes – these show where text on this page has matched part of the Match Expression
* Below the bottom of the page is shown the best matching date that has been retrieved from the document.
* Again below the page at the bottom right corner is shown whether the document matched / failed (the box will be green or red). The number in brackets is from the Match Factor parts of the Match Expression if there are any Match Factors. The number is the sum of the Match Factors for Text Elements that were matched in the document.
* As the mouse is moved over the page the text that has been recognised from that location is displayed
* Buttons for moving to other pages in the document (Next / Prev)

All of the information above will be changed immediately any changes are made to the Match Expression and Date Locations boxes. So it can be quite quick to check out any ideas for improving the document matching or date finding by just making the changes and seeing what comes up on the page. You can also select other documents in the list of results and instantly see whether that document would match or not.

# Document Types

The aim is to get each kind of document (bank statement, electricity bill, etc) to be recognised correctly so that filing can be as simple as possible. To do this each kind of document is defined by a Document Type. The Document Type contains the following information which helps with the process of recognising and filing the appropriate type of document:

* Name (of the Document Type) – this should be something specific like HSBC Statement for 41378791
* Match expression - this contains a combination of text that is uniquely found in this kind of document
* Date locations – defines where to find dates on a document of the this kind and any text that might be close to the right date in the document
* Move-To Folder – this is where documents of this kind should be filed
* Rename-To – this is how to name the document when it is filed
* Enabled – Document Types can be disabled if needed – this can be useful when working on a new document type before it is ready to be used fully
* Thumbnail image (a small picture of a document page) – in cases where a kind of document is hard or impossible to recognise from its text (e.g. a drawing or photo) the Document Type can contain a thumbnail. When filing documents the user can then select the right Document Type by selecting for a list of thumbnail buttons.

## Match Expressions

Match Expressions can be quite complex (it cannot be empty – empty simply means it won’t match any document). The main elements of a Match Expression are as follows:

* Text Elements – any plain text (letters, numbers, spaces) in a Match Expression represents some text that the program will try to find in the document. Matches are done ignoring case – i.e. BanAnA will match banana and BANANA.
* Boolean Operators – these perform the operations AND &, OR | and NOT !
* Round Brackets – the ( and ) can be used to group Text elements and Boolean operators to make it clear what is intended
* Location Rectangles – these follow a Text Element and indicate where the Text Element must be found on the document pages – they are a list of four numbers separated by commas and surrounded by curly brackets { and } – to enter one in Edit Mode click and drag on a document image
* Match Factors – these are used to help indicate which Text Elements are most specific to a type of document. They follow a Text Element and are a colon : followed by a number.

## Match Expressions Example 1

## aviva & ( A10557660J | NR1 3WG)

A document which matches this Match Expression will contain:

The word aviva (or Aviva, AVIVA, aViVa, etc)

AND

Either the text A1055766J OR NR1 3WG

The brackets ensure that it is clear that only one or other of A1055766J, NR1 3WG are needed.

## Match Expressions Example 2 – With Location Rectangle

## HSBC & Statement & 5434{24,34,45,15} & ( 0826|3220|1870|3695)

In this example the match will only happen if the document contains:

The words HSBC AND Statement

AND

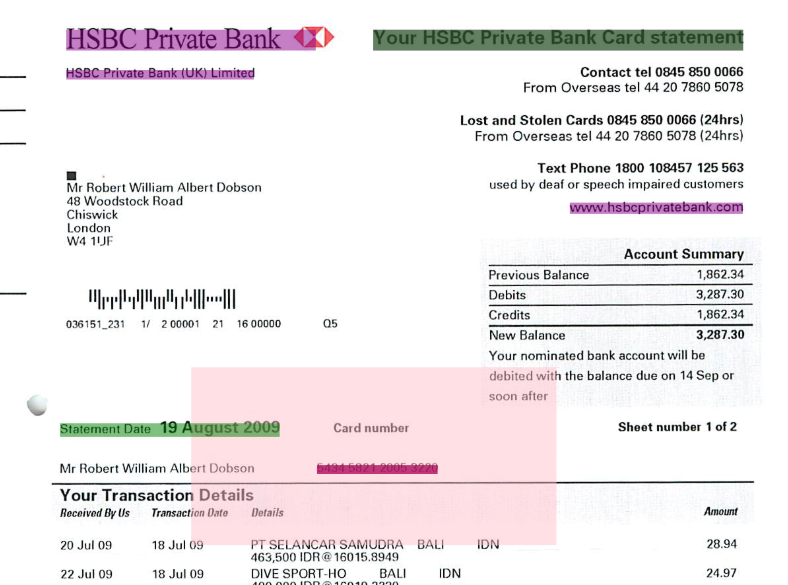
The text 5434 in a location defined by the rectangle 24,34,45,15 (left, top, width, height)

AND

The text 0826 OR 3220 OR 1870 OR 3695

It’s useful to know that this Match Expression was designed to recognise HSBC credit card statements for four different card numbers which all started 5434 and ended 0826 or 3220, etc. So the card numbers were something like 5434 XXXX XXXX 3220, etc.

This is part of a page from a statement that is matched:



The pink coloured box is the one representing the rectangle in which the text 5434 must be found for the document to match.

The other coloured boxes show where the Match Expression parts match text in the document. So, for instance the top two purple boxes and the purple one on the right side all match the text HSBC.

## Match Expressions Example 3 – With Match Factor

## actix:20 & Hammersmith:40

In this (contrived) example the Match Expression will match a document which contains the words:

Actix

AND

Hammersmith

The Match Factors are the numbers in orange and must have a colon preceding them to be recognised as Match Factors. What happens is that a document which contains just actix would get a Total Match Factor of 20, one which contains just hammersmith would get a Total Match Factor of 40 and one which contains both actix and hammersmith would get a Total Match Factor of 60 (which is the case for the document shown below – the Total Match Factor is shown in brackets after the word MATCHES in the green box.



This is useful because the method for matching a document checks through every different possible Document Type and ranks the results based firstly on whether the document matches the Match Expression but secondly on its Total Match Factor.

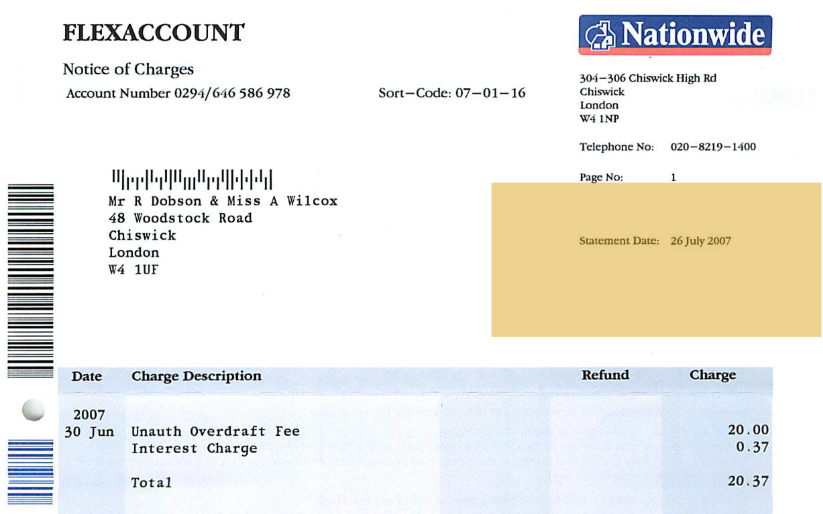
So it is possible to have a Document Type which might both be hard to define absolutely reliably – e.g. a magazine article – but which might be identified fairly reliably by using Match Factors on words that regularly appear in that kind of magazine article – for instance many health and beauty articles might contain the “make-up” but it would not be very reliable to say that this definitely indicates a document of that type.

So by giving the word “make-up” a Match Factor it would mean that this kind of document might be recognised in cases when no other Document Type matches with a higher Total Match Factor.

## Date Locations

Unlike Match Expressions the Date Locations box can be empty. If it is anything at all that is recognised as a date in the first few pages of the document might be used as the document date. Generally the first date found is the most likely to be selected to be used.

Sometimes it helps to define a location for the date for a particular Document Type. For example here is a document that needs a date location to get the date right:



The problem is that in this case the Sort-Code for the account (07-01-16) looks a bit like a date and the program pulls that out as the suggested date for the document because it appears before the correct date which is lower down on the right. As can be seen a mustard coloured box has been placed around the area where the right date resides. This is done by entering Edit Mode and then clicking in the Date Locations box to put the cursor there. Then click and drag on the document view to create (or change) the rectangle in which the date must be found. Once this is done the date appears in the box in the same format used for Match Expressions i.e. something like this:

{59,17,40,13}

More sophisticated Date Locations can include a Match Factor (like the Match Factor in a Match Expression), a Text Element (which must be found before the date in the same text string in the file).

Multiple Text Locations can also be defined.

# Thumbnails

To set the thumbnail for the current Document Type you need to be in Edit Mode and press the USE IMG button. The thumbnail is shown just below this button. To clear the thumbnail press the CLEAR button next to it.

# Notes and Recommendations

1. The Match Expression is the most important part of the Document Type to get right. The expression should match all the documents of the required type and no other documents – i.e. there should be no false-positives. To achieve this in all cases is tricky if not impossible but the better the Match Expression is the faster and more accurate filing can be.
   1. Think of text elements that uniquely occur in the document to be recognised. For instance account numbers, post codes, VAT numbers, telephone numbers, etc. can be very good text elements to choose.
   2. Experiment with the FIND MATCHES button to check against all previously filed (and unfiled) documents to ensure your Match Expression is as good as you can make it.
   3. Use the Match Factor where needed to give more weight to some text elements than others when it is hard to define a document type completely perfectly.
   4. Use the Match Location rectangles where needed to define where on a document the matching text elements should occur.
2. Thumbnails should only be given to document types that can’t be recognised from their text so that the user is only presented with a small number of thumbnails when filing an unrecognised document – fewer choices will make it faster to find the right thumbnail.
3. Note that there should be only one Document Type with any particular name – don’t create a second Document Type with the same name as another as this will lead to confusion.
4. Ideally you shouldn’t assume that all text will be recognised in the scanned document completely correctly. So it’s best to use relatively short pieces of text as there’s a greater probability of an incorrect character in a longer run of text.
5. The software relies on an algorithm to decide how big a gap between two letters to accept as part of the same sentence. Sometimes it gets this wrong so it isn’t a good idea to make too much use of long sequences of words as they might get broken up into smaller sequences and then they won’t match against a long sequence.
6. Unfortunately the set of documents that have already filed have not always been filed correctly. When working through documents that match/don’t-match existing filed documents it is possible that misfiled documents will be arrived at. It is important to make a note of these misfiled documents – at least record the “Unique Name” and any relevant details about the misfiling so that we can later correct these errors and/or remove the document from the reference list.
7. Around 7000 documents were filed with the previous version of this software and the results of filing these documents has been loaded into the new software to enable them to be used for testing new Document Types. In a small number of cases some part of the old record – such as the location of the thumbnail file – might be missing so please don’t assume it perfect.