Orientações para Digitalização

Estas orientações se referem aos critérios adotados no TRF4.

As configurações podem ser utilizadas para qualquer software.

Resumo do conteúdo:

- Digitalizações devem ser feitas em monocromático, 300 dpi (no máximo), com opção de OCR ativada. Se o original estiver ótimo, 200, 240 dpi é suficiente.
- Notas fiscais, fotos e documentos coloridos em geral devem ser digitalizados em 100 dpi. Não há razão para documentos a serem visualizados pelo monitor receberem configuração superior.
- Deve-se privilegiar o formato pdf.
- Por que devo usar OCR? Se o documento é digitalizado sem OCR, ele está igual a uma foto: não se pode selecionar o texto e copiar e, principalmente, não é possível INDEXÁ-LO. O que vem a ser a indexação? O texto é lido pelo programa e seu conteúdo reconhecido e armazenado. Quando preciso localizá-lo, utilizo um programa de busca (como a caixa de pesquisa do SEI) e ele vai até o banco de dados comparar o que escrevi com o que está guardado. É o OCR que permite este registro no banco de dados, dos documentos que digitalizei.

Teoria

Um original de qualidade regular em diante não se beneficiará de uma digitalização mais alta, já que o objetivo é visualização no computador, e não sua impressão.

A resolução de uma imagem digital é a sua definição.

As imagens nas telas são formadas pela justaposição de pequenos pontos quadriculados, chamados "pixels".

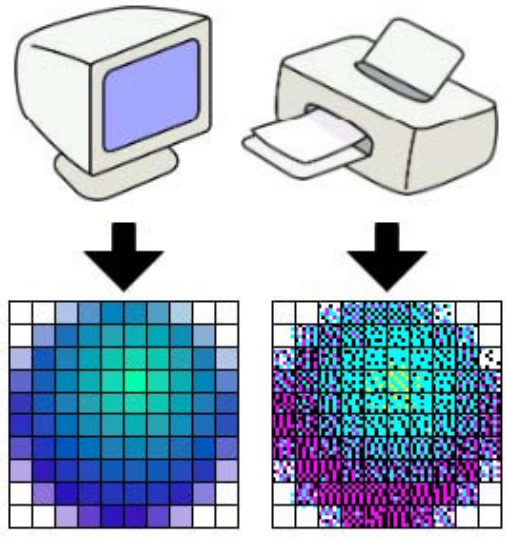
A resolução é medida pela quantidade de pixels na imagem. Sua unidade de medida é o "ppi", que significa "pixels per inch" ou pixels por polegada.

A "dpi" - "dots per inches" diz respeito à quantidade de pontos para uma impressão de qualidade, por isso, em termos fotográficos diz-se "ppi", que traduz a quantidade de pixels por linha do sensor ou da ampliação da fotografia.

Nas imagens, quando maior sua resolução, mais pixels haverá por polegada em altura e largura : assim, imagens de "alta resolução" possuem "pixels" pequeninos, até mesmo invisíveis a olho nu, e, imagens de "baixa resolução" possuem "pixels" grandes que acabam por dar o efeito "pixelation", que deixa imagem quadriculada pelo o tamanho exagerado de seus pontos. Isso é comum acontecer quando tentamos ampliar uma imagem de "baixa resolução"..
Se uma página tiver muitas palavras não reconhecidas ou um texto muito pequeno (abaixo de 9 pontos), tente digitalizar a uma resolução maior.

Digitalize em branco e preto (monocromático) sempre que possível.
Textos abaixo de 9 pontos (fonte menor que 9) o OCR não reconhece, **independentemente da qualidade de digitalização.**

A imagem abaixo demonstra a diferença da mesma imagem obtida pelo monitor e pela impressora, daí a diferença entre pixel e dot.



À esquerda a imagem em PIXEL POR POLEGADA e à direita, PONTO (DOT) POR POLEGADA. Nem sempre uma imagem que aparece boa na tela terá uma impressão da mesma qualidade.
Da mesma forma, se uma imagem será para visualização exclusiva pelo monitor, sua digitalização em muitos pontos não alterarão o resultado da visualização base.
Para a maioria das páginas, a digitalização em preto e branco a 300 ppi produz o texto mais adequado para conversão. A 150 ppi, a precisão do OCR é levemente mais baixa e ocorrem mais erros de reconhecimento de fontes; em uma resolução de 400 ppi ou mais, o processamento fica mais lento e as páginas compactadas são maiores.

Outros dados de digitalização

COMPACTAÇÃO selecione MONOCROMÁTICA CCITT Grupo 4.

Por que? As imagens ocupam arquivos muito grandes. Por isso foram criados ALGORITMOS DE COMPRESSÃO.

A compressão de dados é reduz o espaço ocupado pelos dados num determinado dispositivo. Comprimir dados destina-se também a retirar a redundância, uma vez que muitos dados possuem informações que se repetem ou que podem ser eliminadas sem perda de qualidade ou de informação. Grosseiramente, podemos dizer, se você tem num arquivo a sequência AAAAAA o algoritmo modificará essa sequência para redundância (em A6), os dados são comprimidos pelos mais diversos motivos. Entre os mais conhecidos estão economizar espaço em dispositivos de armazenamento, como discos rígidos, ou ganhar desempenho (diminuir tempo) em transmissões.

O que é o CCITT?

CCITT (Comité Consultatif Internationale de Telegraphie et Telephonie) é um sistema de compressão usado para imagens criado para as transmissões via fax. Para que as transmissões de imagem fossem mais rápidas, criou-se este sistema.

E grupo 4? É o mais recente.

Colorido/Tons de Cinza: Marcar SEM PERDAS.

Por que? Esta é a forma mais conhecida de se classificar os métodos de compressão de dados. Diz-se que um método de compressão é sem perdas (se os dados obtidos após a descompressão são idênticos aos dados que se tinha antes da compressão. Esses métodos são úteis para dados que são obtidos diretamente por meios digitais, como textos, programas de computador, planilhas eletrônicas, etc., onde uma pequena perda de dados acarreta o não funcionamento ou torna os dados incompreensíveis. Um texto com letras trocadas, uma planilha com valores faltantes ou inexatos, ou um programa de computador com comandos inválidos são coisas que não desejamos e que podem causar transtornos. Algumas imagens e sons precisam ser reproduzidos de forma exata, como imagens e gravações para perícias, impressões digitais, etc.
Por outro lado, algumas situações permitem que perdas de dados poucos significativos ocorram. Em geral quando digitalizamos informações que normalmente existem de forma analógica, como fotografias, sons e filmes, podemos considerar algumas perdas que não seriam percebidas pelo olho ou ouvido humano. Sons de frequências muito altas ou muito baixas que os humanos não ouvem, detalhes muito sutis como a diferença de cor entre duas folhas de uma árvore, movimentos muito rápidos que não conseguimos acompanhar num que eles não estão lá. Nesses casos, podemos comprimir os dados simplesmente por omitir tais detalhes. Assim, os dados obtidos após a descompressão não são idênticos aos originais, pois "perderam" as informações irrelevantes, e dizemos então que é um método de compressão com perdas .

Um exemplo bem popular de compressão com perdas é o MP3, pois são eliminadas freqüências inaudíveis ao ouvido humano.

Outras opções

OCR (o que é)

Quando um scanner lê a imagem de um documento, ele converte os elementos escuros da página em um mapa de bits, esse mapa é uma matriz de pixels quadrados que podem estar ativos (pretos) ou inativos (brancos).

O programa de OCR lê o bit gerado pelo scanner e verifica as áreas de pixels ativos e inativos da página, na realidade ele marca o espaço em branco da página.
O Acrobat executa essa operação criando uma camada invisível de texto sobre o documento, na qual só existem as matrizes de pixels ativos, ou seja, só as letras. Como se um acetato transparente fosse colocado sobre a folha e inscritas as letras reconhecidas pelo software.

O espaço em branco entre as linhas de texto contidos em um bloco define a base de cada linha, um detalhe essencial para o reconhecimento de caracteres num texto.
Na primeira etapa de conversão de imagens em texto, o programa tenta reconhecer cada caractere através de uma comparação pixel a pixel com o modelo de letra que o programa guarda na memória.
Os caracteres não reconhecidos passam por um processo mais minucioso e demorado conhecido como extração de recursos, o programa calcula a altura x do texto, relativa à altura da letra minúscula x, e analisa cada combinação das linhas retas, curvas e áreas preenchidas de cada caractere, como no caso da letra o ou da b.

Os programas OCR sabem, por exemplo, que o caractere com uma curva Como o programa elabora um alfabeto de trabalho de cada novo caractere encontrado, a velocidade de reconhecimento aumenta.
Alguns programas OCR marcam os caracteres não reconhecidos com um caractere especial como ~, #, ou @ e desistem.
Outros programas de OCR ainda solicitam um corretor ortográfico especial para procurar erros óbvios e localizar as possíveis alternativas para as palavras que contêm caracteres especiais não reconhecidos.
Por exemplo, para os programas de OCR, o número 1 e a letra I são muito parecidas, da mesma forma que o 5 e o S, ou ainda o cl e o d.

Uma palavra como aclimatar poderia transformar-se em adimatar, o corretor ortográfico reconhece esses erros típicos do OCR e os corrige.

A maioria dos programas de OCR permite que um documento convertido seja gravado em ASCII ou em um formato possível de ser reconhecido pelos processadores de texto e planilhas eletrônicas mais conhecidas.

OUTRAS DIGITALIZAÇÕES

Ao máximo possível, deve se evitar digitalização colorida. Caso seja indispensável, sugere-se utilizar a configuração de 100 dpi. Lembrando que essa digitalização é para que os documentos sejam visualizados na tela e não impressos.