ORIENTAÇÕES PARA DIGITALIZAÇÃO DE DOCUMENTOS

Regras gerais:

- As digitalizações devem ser feitas em monocromático, 300 dpi (no máximo), com opção de OCR ativada. Se o original tem qualidade ótima, 200 ou 240 dpi já são suficientes. **Digitalize em branco e preto (monocromático) sempre que possível.**
- 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.

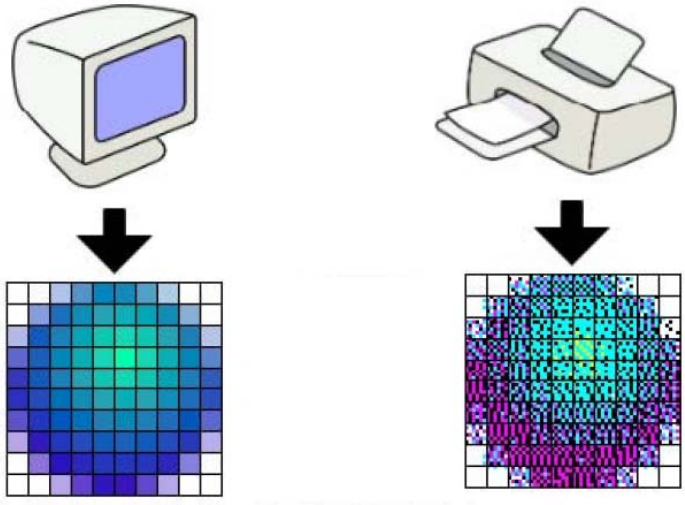
Por que devo usar OCR?
Se o documento é digitalizado sem OCR, ele está igual a uma foto: não se pode selecionar o texto e copiar e, principalmente, não é possível INDEXÁ-LO. O OCR é um programa que permite identificar o texto contido numa imagem. Programas de OCR mais eficientes permitem identificar inclusive textos manuscritos.

O que vem a ser a indexação?
O texto é interpretado pelo Sistema, via OCR, e seu conteúdo reconhecido e armazenado. Se for necessário localizá-lo, utiliza-se a caixa de pesquisa do SEI. O SEI-MP pesquisa o banco de dados e recupera as informações que “batem” com o que foi digitado na caixa de pesquisa.

.

Há benefício real ao aumentar a qualidade da digitalização
Não. Um original com qualidade regular ou acima não se beneficia de uma digitalização com qualidade mais alta, já que o principal objetivo é visualização no computador, e não sua impressão. As imagens nas telas do computador 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 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. 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".

Porém, quando uma página tiver muitas palavras não reconhecidas ou um texto muito pequeno (abaixo de 9 pontos), ai sim vale a pena digitalizar a uma resolução maior.

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 alterará o resultado da visualização base.

Para a maioria dos documentos, 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 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 “sem perdas” deve impossibilitar perdas. Os dados obtidos após a descompressão DEVEM SER idênticos aos dados que se tinha antes.

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, são aceitáveis algumas perdas que não são

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 ou movimentos muito rápidos que não conseguimos acompanhar. Nesses casos, podemos comprimir os dados com um método de compressão “com perdas”.
Um exemplo bem popular de compressão com perdas é o MP3, pois são eliminadas frequências inaudíveis ao ouvido humano.