Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informática Agropecuária
Ministério da Agricultura e do Abastecimento

Instruções Técnicas da Embrapa Informática Agropecuária

Nº. 1, dezembro/2000

Ministério da Agricultura e do Abastecimento

PROCEDIMENTOS PARA DIGITALIZAÇÃO E CONVERSÃO DE DOCUMENTOS PARA O FORMATO PDF

Suzilei Francisca de Almeida Gomes Carneiro[1]

Termos para indexação: PDF; Digitalização; Documento eletrônico.
Index terms: PDF; Digitalization; Electronic document.

1. Introdução

O Projeto Agência é uma aplicação para o ambiente Internet que tem como objetivo divulgar e tornar disponível a informação gerada pelas pesquisas da Embrapa.

Para tanto, é necessário que essa informação que hoje se encontra em meios diversificados seja transformada para o formato eletrônico. O PDF *Portable Document Format* é o formato predominante na Internet hoje para distribuição de informação eletrônica. Ele permite que o texto seja distribuído exatamente como no original em relação à pagina, títulos, gráficos e figuras.

Este documento objetiva auxiliar o usuário a criar ou converter um documento para o formato PDF, através do Adobe Acrobat versão 3.0.

2. O formato PDF

O PDF é hoje um padrão mundial de visualização para distribuição de documentos eletrônicos de propriedade da *Adobe Systems Incoporated*. É um formato de arquivo que preserva o *layout* original em relação as fontes, a formatação, as cores e os elementos gráficos de qualquer documento de origem, não importando qual aplicativo ou plataforma foi usada para criá-lo. É como se cada página de um livro fosse fotografado e suas fotos fossem agrupadas em um só arquivo digital, com facilidades para navegação.

É possível converter qualquer documento em PDF. É um formato que permite compactar arquivos ao serem convertidos e, através do *software* Acrobat Reader, podem ser compartilhados, visualizados e navegados no *browser*, e impressos exatamente como o original.

A criação dos arquivos no formato PDF pode ser realizada através de captura ótica, mediante a utilização de um *scanner*, ou através dos *softwares* mais comuns do mercado que envia os dados da aplicação para impressora, os quais são capturados e transformados em um arquivo PDF, através dos *drivers* PDRWriter e Distiller.

3. Cuidados na pré-digitalização de documentos

O processo de digitalização poderá ser comprometido se o estado de conservação dos documentos apresentar características como:

[1] Informática Agropecuária, Tecnóloga em Construção Civil, Caixa Postal 6041, Barão Geraldo - 13083-970 -

- documentos com páginas amassadas e/ou manchadas;
- documentos deteriorados pelo tempo, mofo ou mesmo de muito manuseio;
- documentos de baixa qualidade gráfica, como por exemplo páginas muito escuras que ao realizar o processo de digitalização o texto do verso seja visível também, prejudicando assim a qualidade do material digital;
- gramatura de papéis muito fina (50 g/m²), também ocorre o mesmo problema anterior.

4. Justificativas para o uso e vantagens oferecidas pelo formato PDF

O Adobe Acrobat, versão 3.0, mesmo sem um verificador ortográfico na língua portuguesa, possui as características : facilidade de criação e publicação de documentos on-line:

- mantém o *layout* original dos documentos digitalizados;
- utiliza o formato de arquivo pdf, que permite a criação de documentos multiplataforma que podem ser visualizados em *browsers* (*software* de navegação na internet);
- possibilidade de captura e conversão de grandes volumes de documentos com um baixo nível de interação do usuário;
- capacidade de compactação do arquivo como texto.
- permite a visualização e navegação por seus arquivos através do *software* acrobat reader, que é um *software* executado dentro de uma janela de um *browser*;
- o pdf permite, também pelo acrobat reader, a realização de pesquisa *full-text* (texto completo);
- o acrobat reader possui ferramentas para auxiliar o usuário na visualização do conteúdo do documento, como *zoom*, movimentação e *layout* das páginas, bem como imprimir um arquivo em partes ou na íntegra;
- pdf ocupa praticamente o mesmo espaço que o formato gif, conforme demonstrado na tabela 1, quando trata a página como imagem, sem proceder o reconhecimento ótico dos caracteres (ocr);
- pdf é voltado para criação e publicação de documentos eletrônicos;
- permite a inclusão de recursos como *hiperlinks, bookmarks, zoom* e outras funcionalidades que facilitam a navegação entre páginas;
- permite a compactação de um documento volumoso em um único arquivo.

TABELA 1. Comparativo entre os tamanhos de arquivo, tomando por base um texto com gráfico, (de 1 página) salvo em vários formatos.

Formato do Arquivo	Espaço Ocupado
BMP Microsoft Windows Bitmap Format	280 Kb
PDF (como imagem)	**74 Kb**
GIF Graphics Interchange Format	74 Kb
JPG Joint Photographi Expert Group	34 Kb
DOC (Word 7.0) - Documentation	39 Kb
PDF (com texto reconhecido)	**21 Kb**
HTML + JPG	71 Kb

O PDF apresenta também vantagens na conversão de documentos tratados como texto e possuindo imagens:

- ocupa aproximadamente 28% do espaço ocupado somente pela imagem;
- permite a realização de pesquisas *full-text* (texto completo);
- permite a seleção do texto podendo este ser transportado para um editor comum com os comandos de copiar e colar.

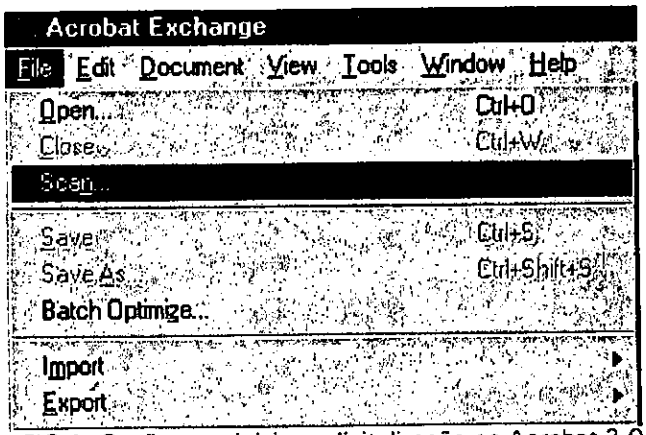

FIG.1. Opção para iniciar a digitalização no Acrobat 3.0.

5. Digitalização de um documento com o Acrobat 3.0

Para iniciar a digitalização de um documento texto usando o Adobe Acrobat versão 3.0, escolha, no módulo Exchange, no menu *File/scan*, como ilustra a Fig. 1.

Em seguida, será aberta uma janela para configuração do *software* do *scanner* que está instalado, no caso da Embrapa Informática Agropecuária é o VistaScan, e clicar em *scan*, como mostra a Fig. 2.

Caso o *scanner* utilizado não tenha um *driver* compatível com o Acrobat 3.0, utilize o *software* do fabricante para digitalizar os documentos e salve-os, preferencialmente, como arquivos TIFF, pois este formato garante que a qualidade do documento digitalizado, inicialmente como imagem, seja preservada.

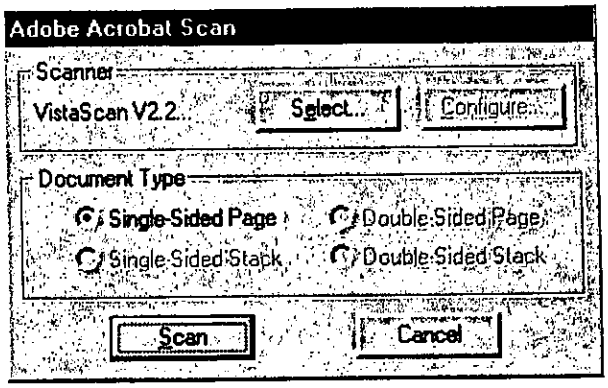

FIG. 2. Caixa de diálogo do *software* do *scanner*.

Os padrões utilizados para digitalização de documentos são o Gray, Halftone e o Lineart .

A diferença entre eles é a intensidade da variações cores entre o branco e preto. O Gray é o padrão utilizado para textos sem gráficos. Possui melhor definição e resolução. O lineart e o halftone são indicados para textos com gráficos, dependendo da qualidade do material a ser escaneado.

É demonstrado na Fig. 3 a interface do *software* ViscaScan32, aberta no momento da digitalização de um texto no Acrobat Exchange, com parâmetros para digitalização.

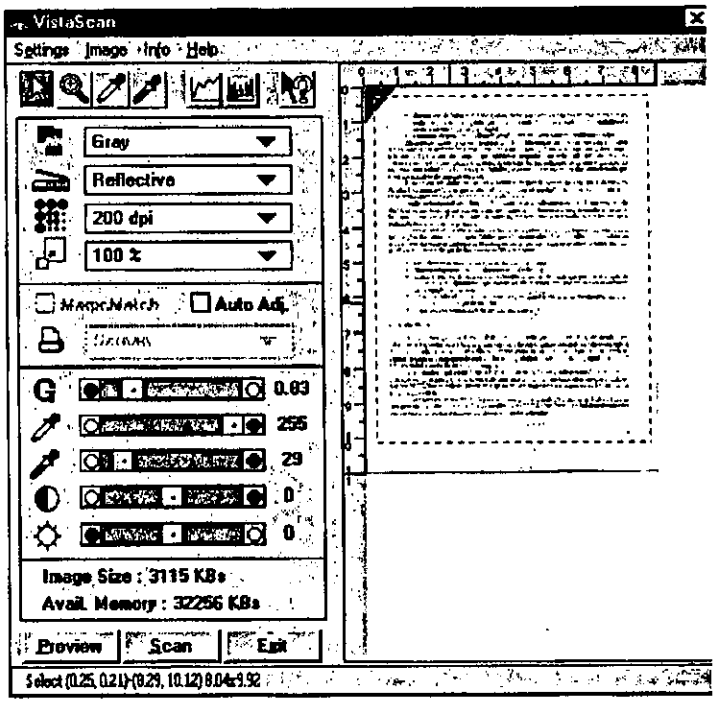

FIG. 3. Caixa de diálogo do *software* VistaScan32 (Twain) com parâmetros para digitalização de um documento (texto).

6. Realização do OCR reconhecimento ótico de caracteres

OCR (reconhecimento óptico de caracteres) é um processo de transformação da imagem em texto através de um *software*. É recomendável que todos os textos no formato PDF estejam habilitados para a realização de pesquisa *full-text* em seu conteúdo.

O *software* Adobe Acrobat versão 3.0 ainda não possui um verificador ortográfico em português, o que o torna inviável na utilização para digitalização de documentos volumosos, por ser um processo muito demorado, quando um documento é digitalizado a partir de um papel. É descrito a seguir os passos para realização do OCR, utilizando o verificador ortográfico em espanhol:

No Adobe Acrobat o processo de OCR no Acrobat é realizado da seguinte forma:

- após o documento ser digitalizado ele será aberto no acrobat exchange;
- configurar o verificador ortográfico para o espanhol, escolhendo na barra de menu *file/preferences/capture;*
- na barra de menu, escolher *document/capture pages*;
- abrirá uma caixa de diálogo do *plug-in* capture, com as opções de leitura de páginas. escolher a que melhor se adapte ao seu trabalho e confirmar com ok;
- o *software* fará o processo de verificação página à página;
- para verificar as palavras que não foram reconhecidas como texto escolher na barra de menu *edit/show capture suspects*. as palavras que ainda permenecem como imagem estarão circuladas em vermelho;
- terminado este processo, é recomendável que o documento seja salvo, se possível, com outro nome, para poder comparar a redução de tamanho de arquivo a partir do original;
- para corrigir as palavras circuladas, escolher na barra de menu *tools/touchup text* ou a letra "t" na barra de ferramenta e clicar sobre qualquer palavra marcada com círculo.
- abrirá, simultaneamente, uma caixa de diálogo com a palavra original e ao clicar sobre a palavra a ser corrigida, esta pode ser editada;
- após a correção, clicar em *accept* na caixa de diálogo e passar para a palavra seguinte;
- é recomendável que o documento seja salvo a cada página tratada ou um conjunto de páginas, para não correr o risco de perder o trabalho realizado;
- é recomendável, também que ao salvar o documento seja escolhida a opção "salvar como" (*save as*) e com a opção otimizada acionada, que garante a compactação do arquivo;
- as palavras que contenham caracteres que o software não reconheça, serão mantidas como imagens, conforme ilustrado na fig. 4.

Quando da aplicação do OCR nas páginas de um documento no formato digital, além da possibilidade de realização de pesquisas, manipulação e edição do texto, estes arquivos ocupam um espaço aproximadamente 4 vezes menor que os arquivos que contêm somente imagens.

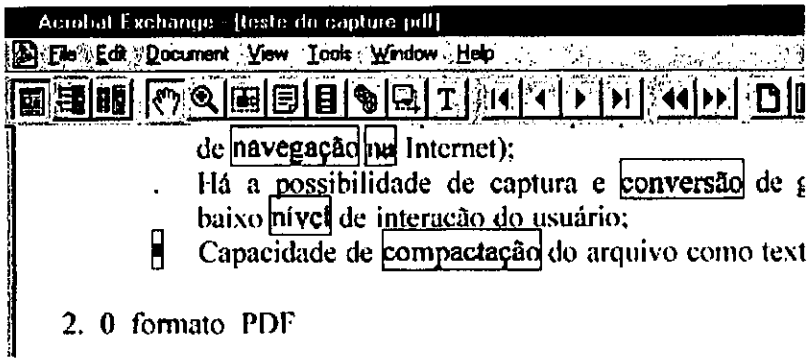

FIG. 4 Palavras não reconhecidas pelo OCR.

7. Parâmetros para manipulação de arquivos PDF

Os *hiperlinks* tornam possível que, através de um clique sobre a referência de um capítulo no índice, se acesse diretamente o conteúdo deste capítulo, ou ainda clicando sobre uma referência bibliográfica, pode-se diretamente ler o documento referenciado na íntegra (se este existir no formato digital). As *bookmarks* são marcadores de páginas que aparecem do lado esquerdo da tela quando o Acrobat Reader está nesta forma de visualização. As *bookmarks* têm função de navegação dentro do próprio documento.

7.1 Inserção de *links* de navegação

Para realizar a inserção dos *links* na página do sumário:

- escolher na barra de menu *tools/link* ou o botão *link* (elo de corrente) na barra de ferramenta;
- clicar e arrastar, fazendo um quadrado ou retângulo sobre a palavra desejada para um *link*;
- abrirá uma caixa de diálogo para configurar o endereço do *link*: *Invisible rectangle/none/go to view/ fit view*;
- antes de criar o *link* através da tecla *set link*, é necessário posicionar na página ou local (destino) para o *link* e só então, confirmar clicando no botão *set link*, conforme ilustra a Fig 5.

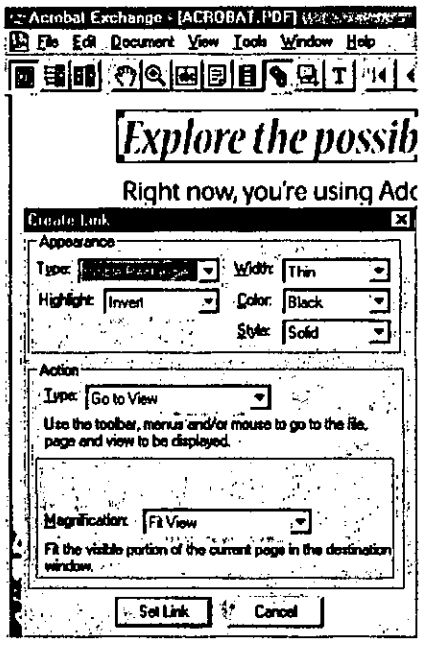

FIG. 5 Configurações do *Link*.

7.2 Inserção de bookmarks (marcadores de páginas)

Para a inserção de *bookmarks*:

- escolher na barra de menu: *document/new bookmark;*
- para definir o destino correspondente da *bookmark* recém-criada, deve-se selecionar a mesma (clicar sobre o ícone criado), na barra de menu e clicar em *edit/properties;*
- abrirá a caixa de diálogo *bookmark properties* para configuração do destino, como ilustra a Fig. 6;
- no campo *type*, define-se o tipo de destino correspondente a *bookmark* que melhor se adapte ao seu trabalho;
- as *bookmarks* permitem que o destino seja um arquivo, uma URL, um som, uma forma de visualização, etc;
- ao clicar sobre a *bookmark* recém-criada, a aplicação alternará imediatamente para o destino especificado na *bookmark*;
- para alterar o nome da *bookmark*, basta clicar sobre ela e dar o nome desejado; é utilizado em documentos volumosos (muitas páginas, como por exemplo um livro digital).

Após selecionado o tipo, a parte inferior desta mesma janela solicitará os dados complementares do destino, como por exemplo, a localização de um arquivo, ou uma URL.

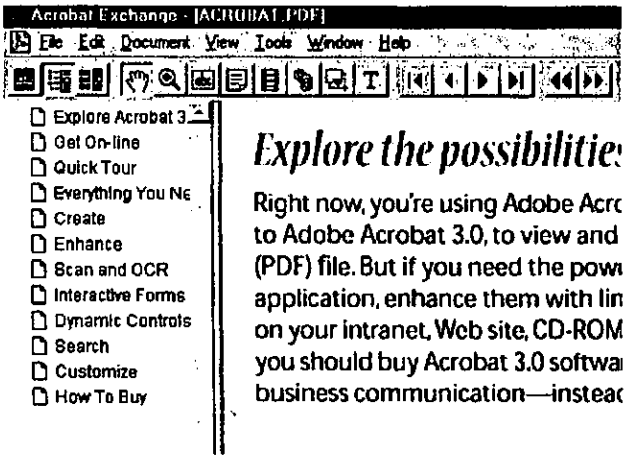

FIG. 6. Visualizando as *bookmarks*.

7.3 Inclusão, exclusão e movimentação das páginas no documento

Uma das características do formato PDF é a manipulação das páginas de seu arquivo. O *software* Adobe Acrobat em sua versão 3.0, permite as alterações:

1. Inserção de páginas (*insert pages*)

- Com o documento que se deseja alterar, clicar em *document/insert pages*;
- abrirá uma caixa de diálogo (*select file to insert*);
- escolher o arquivo a ser inserido e clicar em *select*;
- abrirá outra caixa de diálogo *insert* com as opções de localização e número de páginas;
- escolher a que melhor se adapte ao seu trabalho e confirmar com OK, como ilustra a Fig. 7.

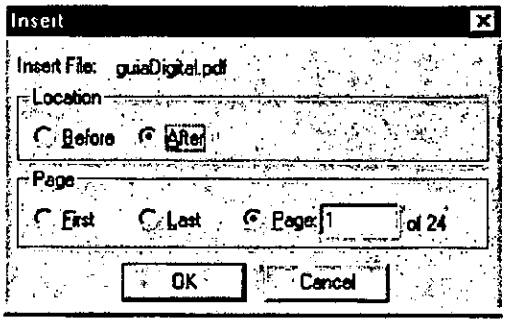

FIG. 7. Inserção de páginas.

2. Extração de páginas (*extract pages*)

- Com o documento que se deseja alterar, clicar em *document/extract pages*;
- abrirá a caixa de diálogo *extract pages* com a opção para digitar o número das páginas a serem extraídas;
- digitar a quantidade desejada e confirmar com OK, como ilustra a Fig. 8.

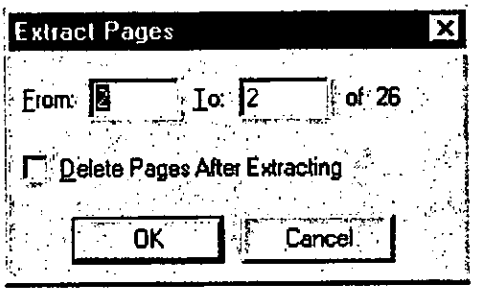

FIG.8. Extração de páginas.

3. Sobreposição de páginas (*replace pages*)

- Abrir o documento na página que se deseja alterar;
- na barra de ferramenta escolher *document/replace pages*;
- abrirá uma caixa de diálogo para escolher o arquivo que irá substituir e clicar em *select*;
- a caixa de diálogo *select file with new page* será aberta
- escolha o arquivo que irá substituir e clicar em *select*;
- abrirá a caixa de diálogo *replace pages* com a opção de localização para substituição da nova página e do arquivo de origem e confirmar com OK, conforme Fig. 9.

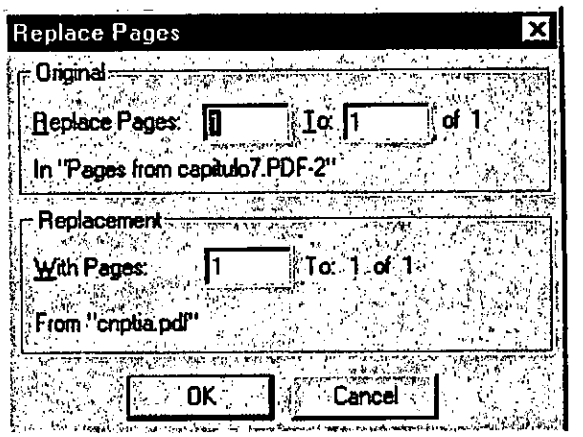

FIG. 9. Sobreposição de páginas.

4. Exclusão de páginas (*delete pages*)

- Com o documento aberto, clicar na barra de menu e escolher *document/delete pages*;
- abrirá a caixa de diálogo com a opção para digitar o número das páginas a serem excluídas;
- digitar a quantidade desejada e confirmar com OK, como ilustra a Fig. 10;
- será aberta uma mensagem de confirmação de exclusão das páginas, pois uma vez excluídas, essas páginas não poderão ser recuperadas.

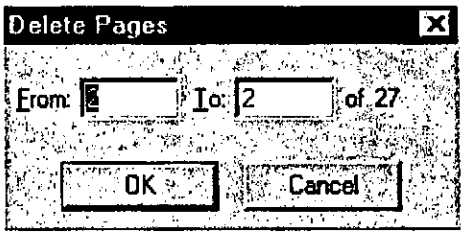

FIG. 10. Exclusão de páginas.

5. Alteração do tamanho das páginas (*crop pages*)

Para alterar o tamanho das páginas utiliza-se, na barra de menu a opção *document/crop pages*.

A seguir, é aberta a caixa de diálogo onde deve-se especificar o tamanho das margens esquerda (*left*), direita (*right*), superior (*top*), inferior (*bottom*), que serão retiradas de cada página, conforme ilustra a Fig. 11.

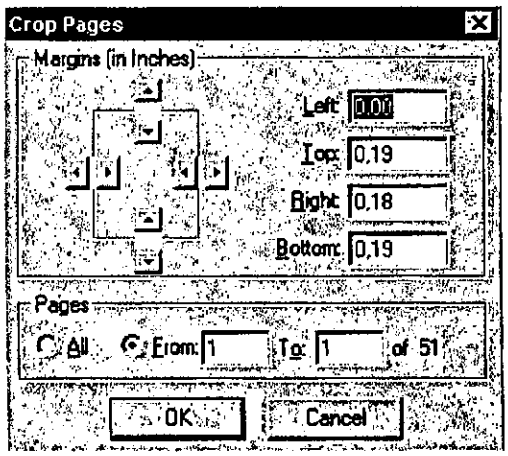

FIG. 11. Definindo o tamanho da página.

Após pressionado o botão OK esta tarefa está terminada.

6. Visualização das páginas

No Acrobat Reader, existem diversas configurações relativas a forma de visualização do texto. Estas opções, descritas na Tabela 2, podem ser encontradas no menu View.

TABELA 2. Configurações de visualização.

Actual Size	Exibe a página em seu tamanho original (100%).
FitPage	Encaixa a página inteira na tela.
FitWidth	Encaixa a largura da página na tela.
FitVisible	Encaixa o texto da página na tela.
Zoom to	Permite que o usuário escolha o valor de zoom da página.
Full Screen	Exibe em tela cheia.
Sigle Page	A transição entre as páginas é descontínua. Ao ultrapassar o final de uma página, a próxima é exibida usando toda a tela.
Continuous	Permite a transição contínua entre as páginas através da barra de rolagem, sendo possível visualizar simultaneamente o final de uma.
Continuous Facing Pages	Transição contínua com as páginas lado a lado (duas páginas na tela).
Page Only	Exibe somente a página na tela.
Bookmark and Page	Exibe a página e as *bookmarks* na parte esquerda da tela.
Thumbnails and Page	Exibe a página e thumnails (miniaturas das páginas) na parte esquerda da tela.

Também é importante proceder a configuração da forma de visualização inicial do arquivo quando este for ser aberto por um usuário em um *browser*.

As opções disponíveis para visualização inicial das páginas no Acrobat Reader devem ser selecionadas considerando a legibilidade das páginas do documento, devendo ser possível realizar a leitura do texto do documento a partir da visualização selecionada.

Portanto, deve-se escolher o conjunto de opções mais adequado para tornar a leitura e navegação do texto agradáveis, bastando testar anteriormente uma configuração adequada e, posteriormente, selecioná-la na janela *Open Info*, ilustrada na Fig. 12, acessada na barra de menu *file/Document Info/Open*.

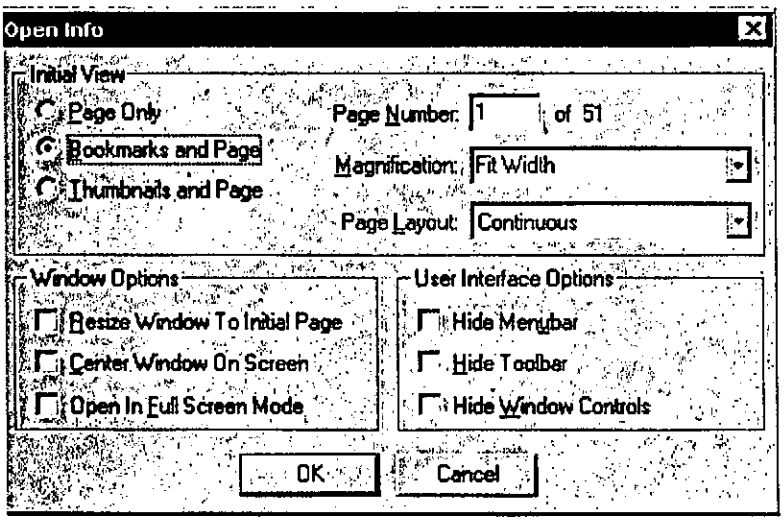

FIG. 12. Configuração recomendada para visualização quando o documento possuir *bookmarks*.

Com isso, o documento digitalizado está preparado para ser acessado, permitindo ao usuário facilidade na leitura do texto e navegação entre as páginas e capítulos do documento.

8. Conversão de documentos através do PDFWriter e PDF Distiller

Para a conversão de arquivos digitais gerados a partir de qualquer *software* que seja compatível com o Windows 95 para o formato PDF, é necessário a utilização do Acrobat PDF Writer ou PDF Distiller.

O PDFWriter e o Distiller funcionam como um *driver* de impressora. Uma vez instalado no Windows eles criam uma nova impressora com os respectivos nomes. Estes *drivers* tornam-se disponível para qualquer *software* do Windows, comportando-se como um *driver* de uma impressora qualquer.

Para que seja criado um arquivo PDF a partir de um *software* qualquer, deve-se ativar o comando de impressão correspondente e, necessariamente, deve ser escolhida a impressora Acrobat PDF Writer ou Distiller, como ilustra a Fig. 13.

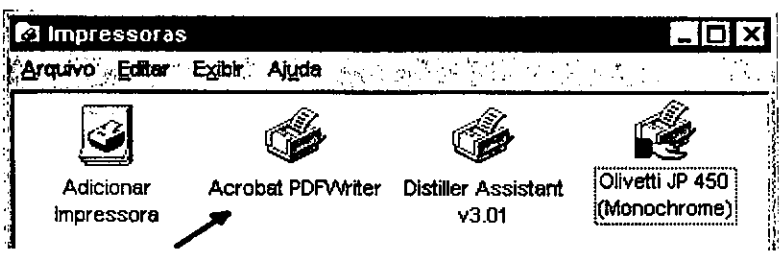

FIG. 13. O *driver* do Acrobat PDFWriter e Distiller.

Ao confirmar o pedido de impressão, o *driver* captura a impressão e solicita um nome para o arquivo PDF que será criado.

Terminada a impressão, o arquivo PDF está pronto para ser visualizado através do Acrobat Reader.

9. Referências bibliográficas

PEOPLE COMPUTAÇÃO. Centro de Treinamento em Informática. **Adobe Acrobat 4.0**. Campinas, 2000. 51p.

ADOBE SYSTEMS INCORPORATED. [**Adobe**: a inspiração transformada em realidade]. [S.l.], 2000.
Disponível em: <http://www.adobe.com.br>. Acesso em 27 jul. 2000.

RAABE, A.; POHLMANN FILHO, O. Estudo comparativo entre sistemas de digitalização de documentos: formatos HTML e PDF. **Ciência da Informação On line**, Brasília, v.27, n.3, 1998.
Disponível em: <http://www.ibict.br/cionline/270398/index.htm>. Acesso em: 17 maio de 2000.

IMPRESSO

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informática Agropecuária
Ministério da Agricultura e do Abastecimento
Rua Dr. André Tosello, s/nº Caixa Postal 6041 - Barão Geraldo
13083-970 - Campinas, SP
Fone (19) 3289-9800 Fax (19) 3289-9594
E-mail: sac@cnptia.embrapa.br
http://www.cnptia.embrapa.br

MINISTÉRIO DA AGRICULTURA E DO ABASTECIMENTO

GOVERNO FEDERAL
Trabalhando em todo o Brasil