Instalação da solução OCR

1. Introdução

Este documento visa a orientação para a instalação da solução OCR a ser utilizada no Instituto Federal de Pernambuco para reconhecimento óptico de caracteres em documentos digitalizados. Tal tecnologia permite a busca, indexação de arquivos PDF, como também assistência a deficientes visuais por meio de softwares assistivos.

Este documento está estruturado como se segue. A seção 2 descreve as características dos softwares utilizados para a solução, na seção 3 demonstra-se os procedimentos de instalação do OCRmyPDF e demais softwares. A seção 4 apresenta as configurações/modificações necessárias para o funcionamento do sistema, incluindo a configuração da impressora caso a TI opte por esta solução. Na seção 5 é demonstrado o uso da solução, e, finalmente, na seção 6 são apresentadas as referências utilizadas.

Características

- Processamento OCR;
- Texto em língua portuguesa do Brasil;
- Arquivos com saída PDF do tipo PDF/A-1b;
- Possibilidade de monitoramento de um ou mais diretórios de entrada;
- Processamento de PDFs mistos (com texto digital e texto escaneado);
- Processamento de documentos grandes (> 50 pag.);
- Manutenção dos arquivos PDF originais;
- Log para registro do processamento da solução OCR.

2. Softwares utilizados

> O sistema operacional recomendado é o **Debian** a partir da **versão 9**.

OCRmyPDF

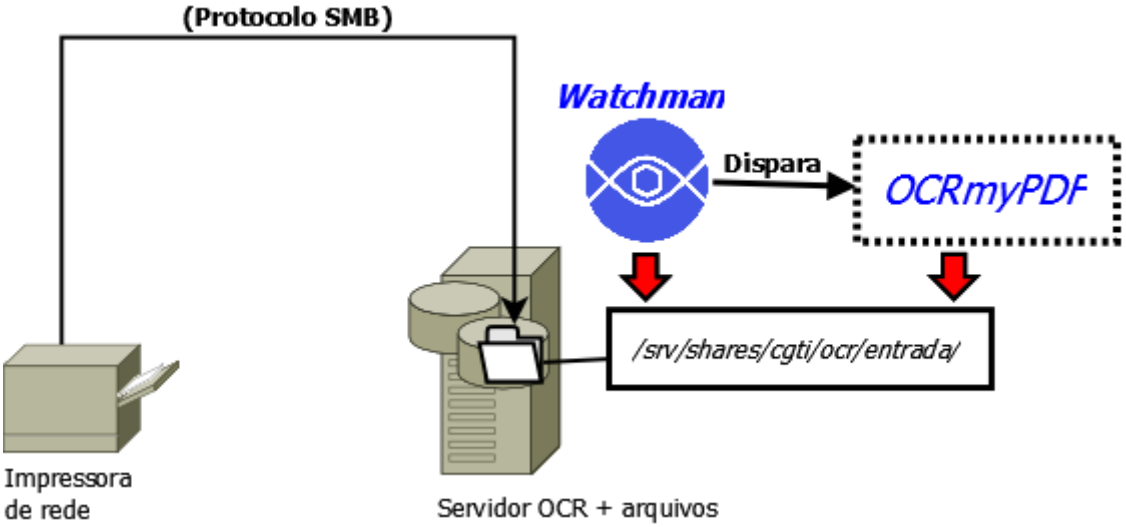
O OCRmyPDF realiza a leitura de documentos escaneados em PDF e permite o reconhecimento do seu texto. Assim, podendo ser buscado e indexado. Além disso, os documentos são automaticamente gerados em formato PDF/A, para arquivamento de longo prazo de documentos eletrônicos. O OCRmyPDF gera documentos no formato PDF/A-1b e PDF/A-2b.

Watchman

O Watchman permite monitorar determinado(s) diretório(s) com o objetivo de realizar operações quando os arquivos são modificados. É possível filtrar os tipos de arquivos com base no nome, tipo, dentre outras características.

Como o OCRmyPDF não funciona com base no esquema de cliente-servidor até o presente momento, é necessário um software como o Watchman para monitorar determinadas entradas e disparar comandos para executar o processamento dos arquivos inseridos ou modificados.

Na imagem seguinte, uma demonstração do fluxo de operação de todo o sistema caso a TI local opte pelo uso de um servidor de arquivos como o Samba.



Na imagem seguinte, uma demonstração do fluxo de operação de todo o sistema caso a TI local opte pelo uso de um servidor de SFTP.

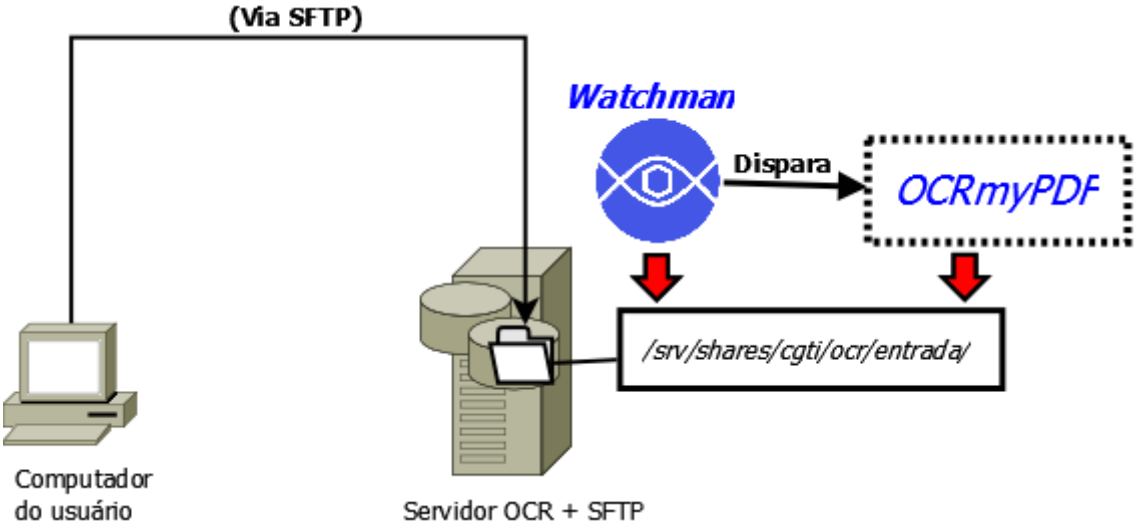


Image not found or type unknown

O método de envio do arquivo para a pasta de entrada não está no escopo desta solução.

Vozes NextUp-ScanSoft

Se desejar, este é um pacote de fala disponibilizado pela NextUp com uma voz feminina chamada Raquel. Este pacote é destinado ao uso nos clientes e não é necessária sua instalação no lado do servidor OCR. Atualmente, o software não faz mais parte do portfólio da empresa, mas a instalação é permitida e sem custos com o instalador disponível na Web. O pacote é utilizado por softwares como Adobe Reader e Foxit PDF Reader para a leitura do texto.

3. Instalação

OCRmyPDF

A última versão do OCRmyPDF disponível nos repositórios Debian/Ubuntu é bem anterior a versão já disponibilizada no repositório oficial dos desenvolvedores. Assim, não contempla as últimas atualizações necessárias ao bom funcionamento do pacote no que se refere a demanda de OCR do IFPE. **Portanto, será demonstrada a instalação via compilação do código fonte.**

Logado como ***root***, execute os comandos:

```
$ apt-get update
$ apt-get install python3-pip
$ apt-get install libffi-dev
$ apt-get install tesseract-ocr
$ apt-get install ghostscript
$ apt-get install qpdf
$ apt-get install git
$ pip3 install git+https://github.com/jbarlow83/OCRmyPDF.git
```

Para dar suporte ao idioma Português do Brasil, instale o pacote abaixo:

```
 $ apt-get install tesseract-ocr-por -y
```

Com o OCRmyPDF instalado, faça o teste com algum arquivo PDF escaneado:

```
$ ocrmypdf entrada.pdf saida.pdf --output-type pdfa-1 -l por
```

Os argumentos `--output-type pdfa-1 -l` por significam respectivamente o tipo de PDF/A gerado e o idioma a ser considerado na interpretação do texto presente no documento PDF.

Watchman

Ainda como ***root***, instale alguns pacotes de dependência:

```
$ apt-get install gcc
$ apt-get install autoconf
$ apt-get install automake
$ apt-get install build-essential
$ apt-get install libtool
$ apt-get install libssl-dev
$ apt-get install pkg-config
$ apt-get install python-dev
$ apt-get install python3-dev
```

Agora procede-se a instalação do Watchman (como ***root***):

```
$ git clone https://github.com/facebook/watchman.git
$ cd watchman
$ git checkout v4.9.0
$ ./autogen.sh
$ ./configure
$ make
$ make install
```

NextUp-ScanSoft Raquel

Apenas baixe e instale o software (Windows) por meio do link abaixo:

```
https://drive.google.com/file/d/0B3aNFZuG_Yw9cjRCNUJBOUQ3QVk/view?usp=sharing
```

Desinstalação

Caso seja necessário remover alguns dos softwares para manutenção/atualização, realize os seguintes comandos à seguir. **Para remoção do OCRmyPDF**:

```
$ pip3 uninstall ocrmypdf
```

Para a remoção do Watchman, seria necessário excluir os binários, portanto, não se recomenda esta prática. Entretanto, caso necessite atualizar o Watchman, basta realizar a nova instalação que os binários serão sobrescritos com a nova versão.

4. Configuração

A configuração para a solução OCR fica praticamente com o Watchman e a impressora, já que a utilização do OCRmyPDF é bem simples. Considere a seguinte estrutura:

```
root@t-ifpe-ocr-beta: /# tree /srv
/srv
└── shares
    ├── ascom
    │   └── ocr
    │       ├── entrada
    │       ├── originais
    │       └── saida
    ├── cgti
    │   └── ocr
    │       ├── entrada
    │       ├── originais
    │       └── saida
    └── dae
        └── ocr
            ├── entrada
            ├── originais
            └── saida
```

Watchman

O Watchman irá monitorar todos os diretórios “entrada” e dispara um script quando um novo arquivo PDF for criado. No exemplo abaixo, iremos criar três triggers. Para cada diretório “entrada”, **crie o diretório e em seguida o seguinte arquivo** de configuração do Watchman, exemplo para **cgti**:

```
$ mkdir -p /srv/shares/cgti/ocr/entrada/
$ mkdir -p /srv/shares/cgti/ocr/saida/
$ mkdir -p /srv/shares/cgti/ocr/originais/

$ nano /srv/shares/cgti/ocr/entrada/.watchmanconfig
{"settle": 10000}
```

> Sobre os diretórios criados, é importante trabalhar corretamente as permissões de usuário e grupo de acordo com o seu serviço de rede e políticas de acesso.

A função do arquivo `.watchmanconfig` é passar configurações específicas do diretório ao watchman. Neste caso, a opção "*settle*" impede que o PDF seja processado antes de sua completa transferência para o diretório. **Crie em cada diretório "entrada" antes de executar o próximo comando** de criação dos triggers.

Os triggers são criados e invocam o script no diretório **/srv/** - o script pode estar em outro local, só necessita de permissão de execução do root. **Crie as triggers**:

```
$ watchman -- trigger /srv/shares/cgti/ocr/entrada/ 'ocrTrigCgti' '*.pdf' -- /srv/doOcr.sh
$ watchman -- trigger /srv/shares/dae/ocr/entrada/ 'ocrTrigDae' '*.pdf' -- /srv/doOcr.sh
$ watchman -- trigger /srv/shares/ascom/ocr/entrada/ 'ocrTrigAscom' '*.pdf' -- /srv/doOcr.sh
```

Detalhes:

- /srv/shares/cgti/ocr/entrada/ : diretório observado;
- 'ocrTrigCgti': nome do trigger;
- '*.pdf': filtro aplicado para o trigger;
- /srv/doOcr.sh: script a ser invocado quando o Watchman identificar alguma alteração no diretório monitorado;
- Importante não esquecer os dois hífens duplos no comando “--”.

Agora observe o script que está sendo utilizado, **copie e cole no mesmo caminho**, ou seja, /srv/doOcr.sh:

```
#!/bin/bash

# Escreve no log. Recebe o nome do arquivo e a ação realizada.
logIt(){
echo -e "$(date +%d/%m/%Y-%T)\t" $1"\t"$2 >> /var/log/ocr.log
}

# Realiza o OCR. Recebe o nome do arquivo de entrada.
ocrFile(){
  logIt "$arg" "Processando"
  filename=$(echo "$1" | cut -f 1 -d '.') # sem extensão
  entrada="$1"
  saida="$filename"-ocr.pdf

  ocrmypdf "$entrada" ../saida/"$saida" -l por --output-type pdfa-1 --skip-text

  chmod g=rw,o= ../saida/"$saida"
  logIt "$arg" "OCR realizado"
```

```
  mv "$arg" ../originais
}

# Fluxo principal
for arg in "$@"
do
    if [ -f "$arg" ]; # arquivo existe ?
    then
       logIt "$arg" "Arquivo recebido"
       ocrFile "$arg" &
    fi
done
```

> Caso copie e cole o script acima, antes, edite o texto e organize as linhas e aspas duplas (“ ”).

Dê permissões de execução para o script.

```
$ chmod +x /srv/doOcr.sh
```

Ativando a auto-inicialização

O watchman deve monitorar os diretórios assim que o S.O. for iniciado. De forma a permitir que esta função seja realizada, **deve-se adicionar o comando de inicialização do watchman** aos comandos executados após o início do S.O.

Para realizar isto, basta seguir o tutorial de scripts de inicialização com system V que está na WIKI do IFPE. Atente para as variáveis do script que neste caso do watchman devem ser:

`dir="/usr/local/bin/"`

`cmd="watchman -f"`

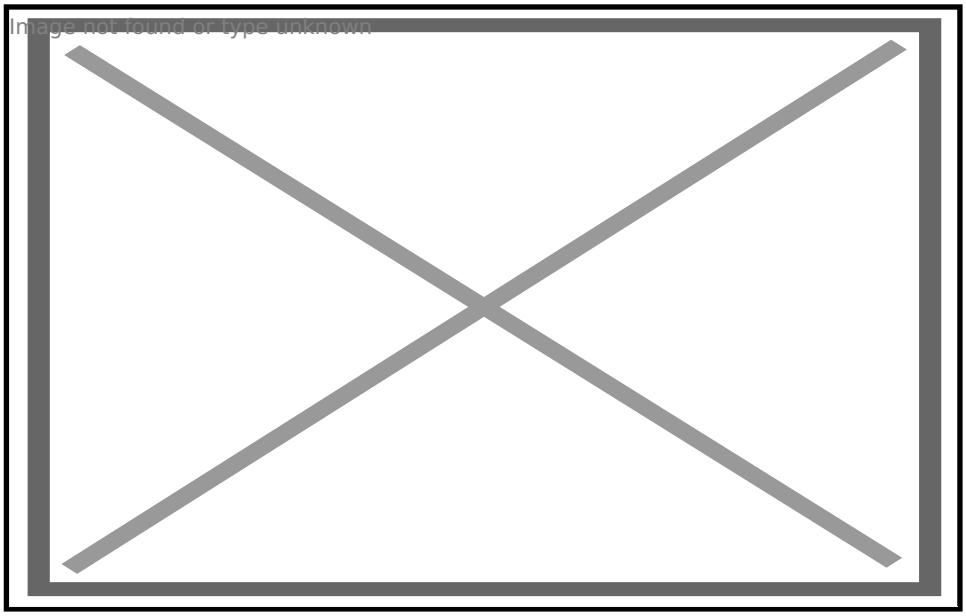
`user="root"`

Impressora

Será demonstrada configuração na impressora Kyocera FS-1135 enviando o PDF para um servidor de arquivos. Para outras impressoras, as configurações são semelhantes, com algumas pequenas diferenças que podem ocorrer na forma como o S.O. da impressora trata os dados inseridos.

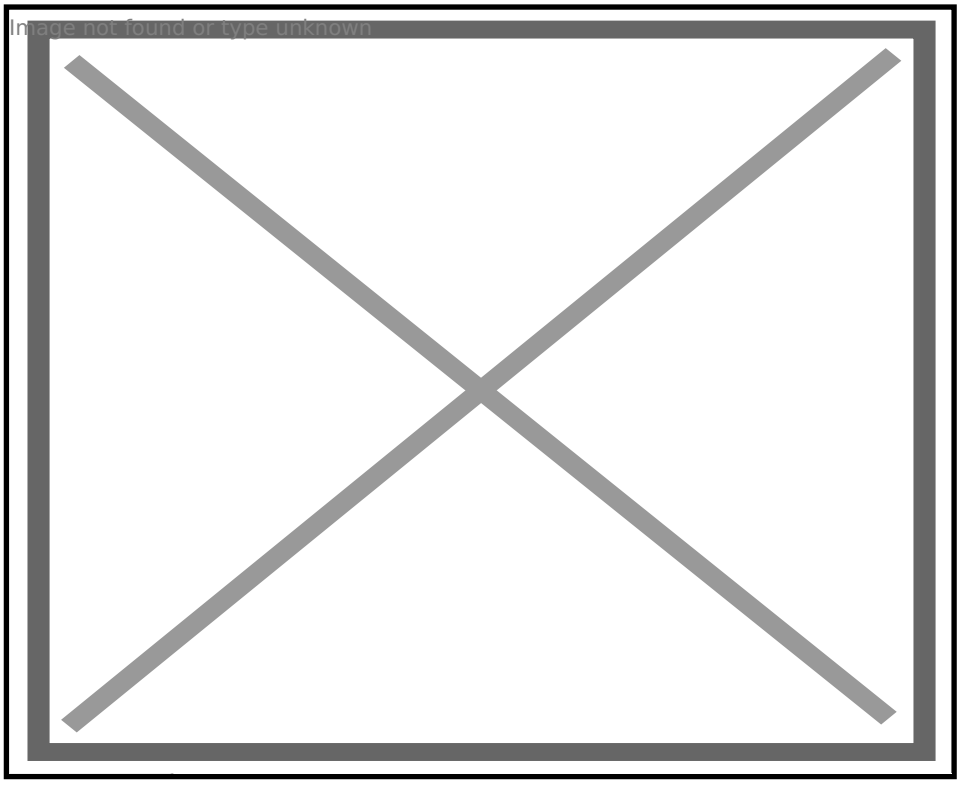
> A impressora deve estar conectada diretamente à rede.

Acesse o painel WEB da impressora:



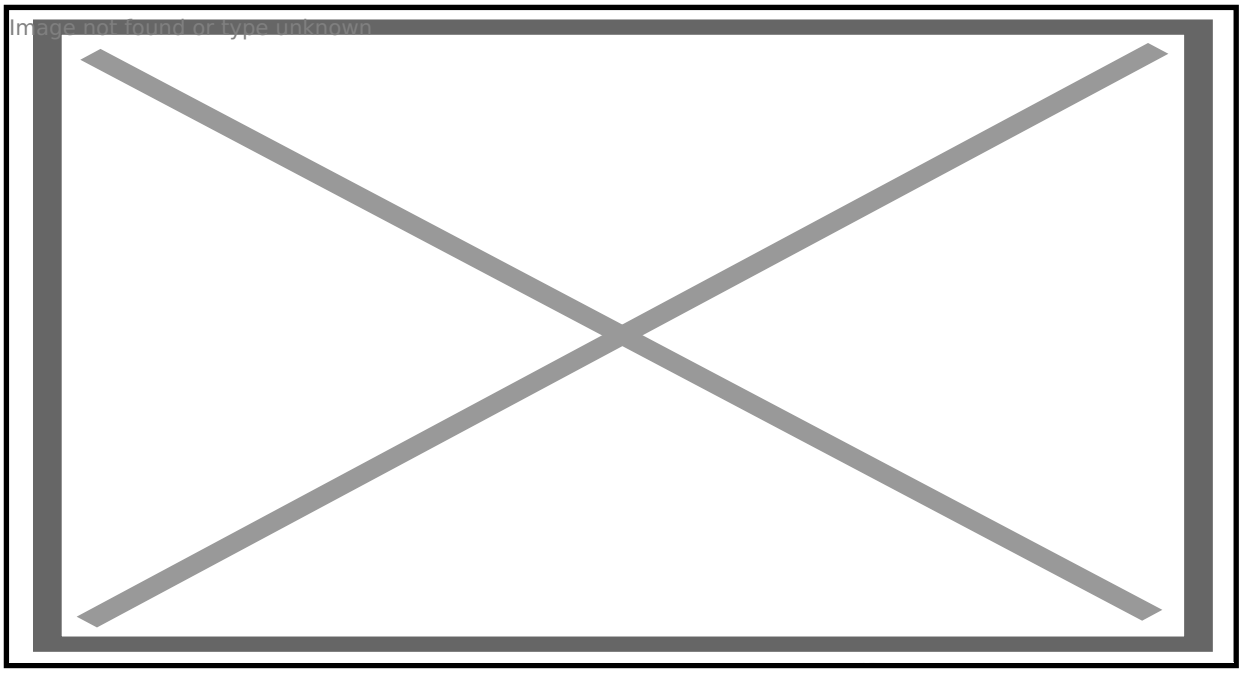
Serão criados “contatos” na agenda da impressora com configurações pré-definidas para salvar os documentos escaneados. Desta forma, evita-se a inserção manual das configurações de acesso ao servidor de arquivos. Na Kyocera, por exemplo, um contato pode conter informações FTP, SMB e de e-mail para o envio do PDF gerado.

Acesse o menu “Básico”, depois, nas opções do lado esquerdo, clique “Bloco de endereços” e em seguida “Contatos”. Crie um novo contato clicando em “Adicionar Contato”.



Tem-se várias opções de envio, será escolhido o protocolo SMB.

Image not found or type unknown



- **Nome de host**: IP ou hostname do servidor de arquivos;
- **Número da porta**: porta que o servidor de arquivos escuta para receber arquivos;
- **Caminho**: utilizando contra-barras, o nome do compartilhamento seguido dos sub-diretórios para onde serão enviados os arquivos PDF;
- **Nome de login**: usuário e DOMÍNIO no formato usuario@dominio. **Recomenda-se a criação de um usuário no seu domínio para cada impressora.**
- **Senha de login**: senha do usuário no DOMÍNIO.

Clicar em **Enviar** para salvar as configurações. Posteriormente, para escanear e mandar para o servidor de arquivos, seguir as próximas configurações.

Envio dos arquivos

A configuração do envio dos arquivos foge do escopo deste tutorial. Mas o importante é que os usuários criados para o envio dos arquivos tenham permissões de **escrita (write)** no diretório “entrada” configurado. **O script (/srv/doOcr.sh) deve ter configurações de execução habilitadas** para que o Watchman consiga dispará-lo, bem como os usuários do domínio devem ter no mínimo acesso de leitura no diretório de “saída”.

Recomenda-se uma política de manutenção dos arquivos que ficarão no diretório originais. Pode-se movê-los para um backup ou executar algum script periódico para limpar arquivos antigos.

LOG

O log de processamento dos arquivos da solução OCR se encontrará disponível no caminho /var/log/ocr.log. O estado possíveis para arquivo é:

- **Arquivo recebido** - o arquivo foi recebido e detectado pelo watchman que já invocou o script doOcr.sh.
- **Processando** - o OCRmyPDF está processando o arquivo.
- **OCR realizado**. - o arquivo foi processado pelo OCRmyPDF com sucesso e movido para o diretório “saida”.

5. Atualização a partir de uma versão anterior

Se já existia uma versão anterior da solução OCR implantada, basta realizar a atualização de alguns pontos.

Atualize o OCRmyPDF:

```
$ apt-get update
$ apt-get install python3-pip
$ apt-get install libffi-dev
$ apt-get install tesseract-ocr
$ apt-get install ghostscript
$ apt-get install qpdf
$ apt-get install git
$ pip3 install git+https://github.com/jbarlow83/OCRmyPDF.git@v5.7.0
$ apt-get install tesseract-ocr-por -y
```

Delete as triggers antigas do watchman, exemplo para **cgti**:

```
$ watchman watch-del /srv/shares/cgti/ocr/entrada
```

Crie o arquivo .watchmanconfig no(s) diretório(s) de entrada do OCR**,** exemplo para **cgti:**

```
$ nano /srv/shares/cgti/ocr/entrada/.watchmanconfig
{"settle": 10000}
```

Faça uma cópia e atualize o script de execução doOcr.sh para esta versão abaixo**:**

```
$ cp /srv/doOcr.sh /srv/doOcr.sh-bkp

$ nano /srv/doOcr.sh



<<<<<<<<<insira o conteúdo abaixo>>>>>>>>>>>>>>>>>>>>>>


#! /bin/bash

# Escreve no log.  Recebe o nome do arquivo e a ação realizada.
logIt(){
echo -e "$(date +%d/%m/%Y-%T)\t" $1"\t"$2 >> /var/log/ocr.log
}

# Realiza o OCR.  Recebe o nome do arquivo de entrada.
ocrFile(){
  logIt "$arg" "Processando"
  filename=$(echo "$1" |  cut -f 1 -d '.') # sem extensão
  entrada="$1"
  saida="$filename"-ocr.pdf

  ocrmypdf "$entrada" ../saida/"$saida" -l por --output-type pdfa-1 --skip-text

  chmod g=rw,o= ../saida/"$saida"
  logIt "$arg" "OCR realizado"
  mv "$arg" ../originais
}

# Fluxo principal
for arg in "$@"
do
    if [ -f "$arg" ]; # arquivo existe ?
    then
       logIt "$arg" "Arquivo recebido"
       ocrFile "$arg" &
    fi
done
```

Crie novamente as triggers do watchman, exemplo para **cgti:**

```
$ watchman -- trigger /srv/shares/cgti/ocr/entrada/ 'ocrTrigCgti' '*.pdf' -- /srv/doOcr.sh
```

6. Operação com scanner

Aqui é demonstrado o envio dos arquivos de uma impressora para o servidor de arquivos local.

Image not found or type unknown

Scanner

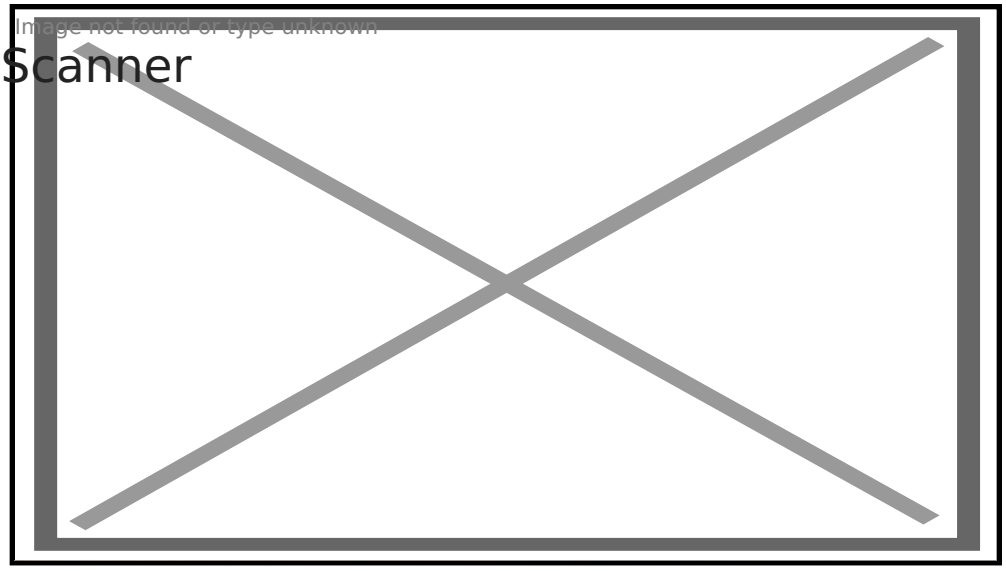
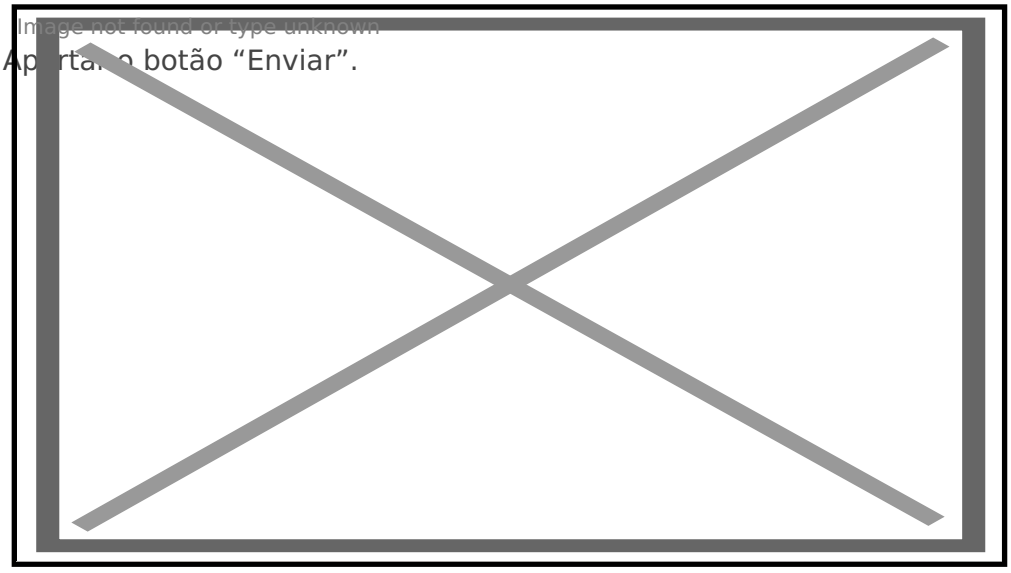


Image not found or type unknown

Apertar o botão “Enviar”.



Apertar o botão de contatos, ou “Libro de direcciones” como neste caso. Tem a função de acesso a Agenda.

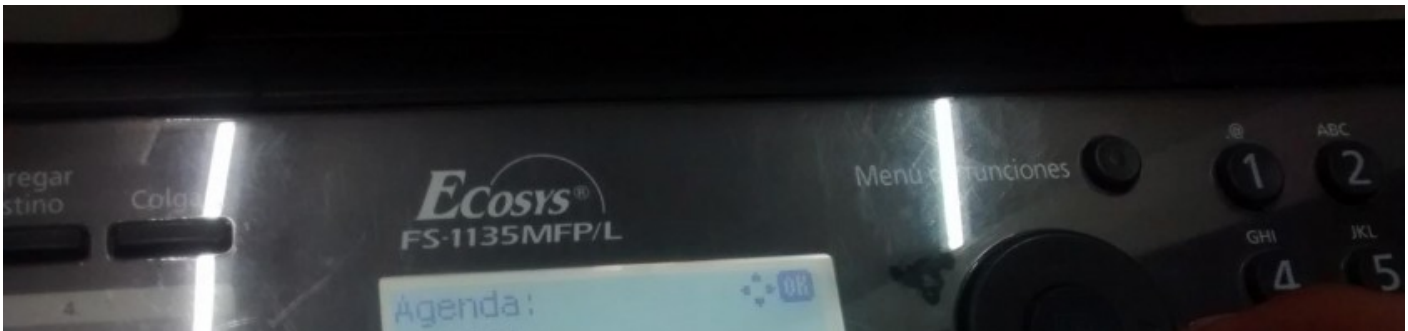


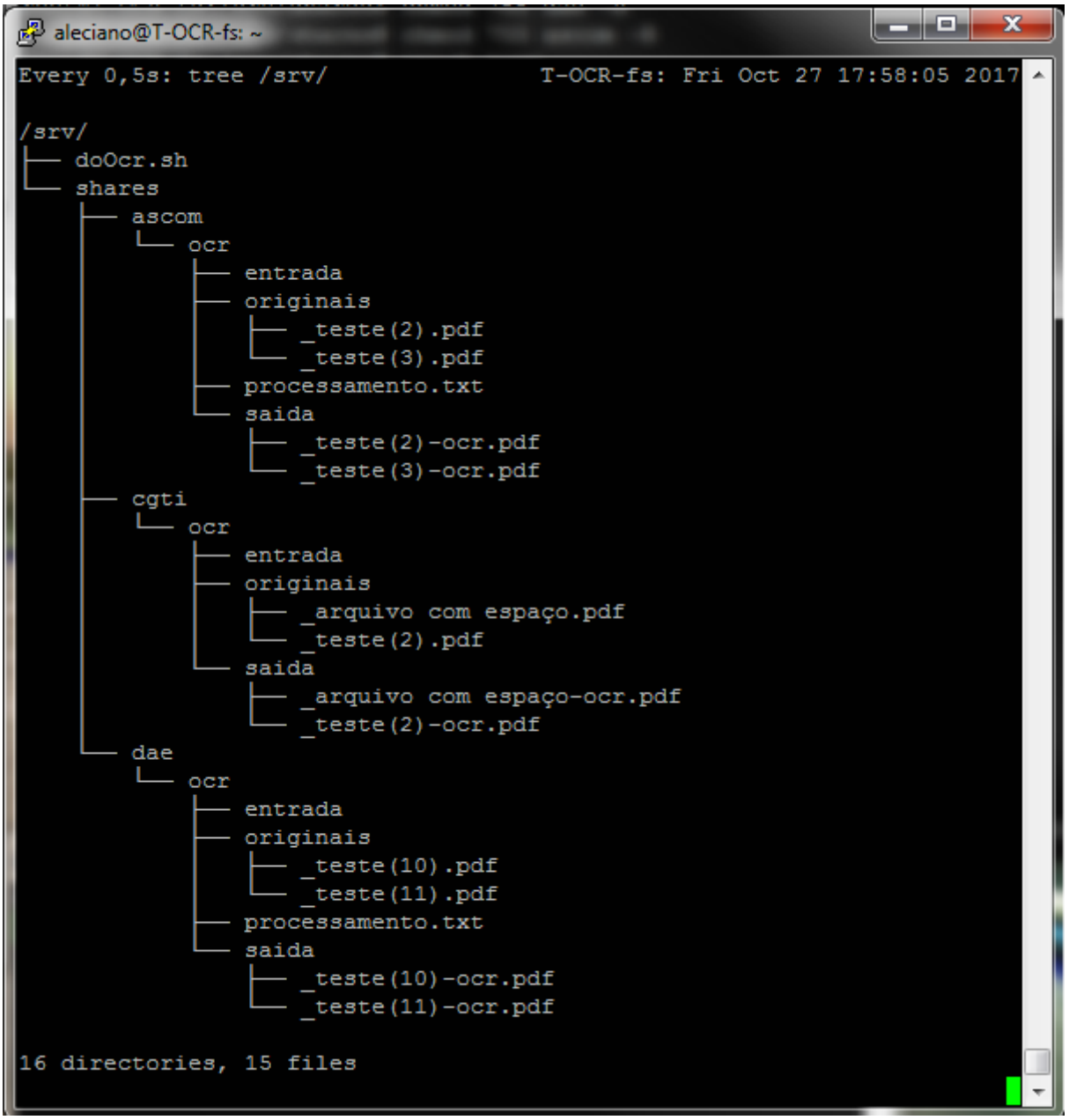
Image not found or type unknown

Selecione o contato previamente salvo na agenda e apertão o botão “Ok”.

Aperte “Ok” para confirmar o endereço. Em seguida, apertar o botão verde principal para dar início a digitalização.

Image not found or type unknown

Se não aparecer nenhuma mensagem de erro ou o LED vermelho não piscar, houve sucesso no escaneamento.



Diretório “Ascom”, na estrutura, indicando sucesso no escaneamento e no processamento OCR.

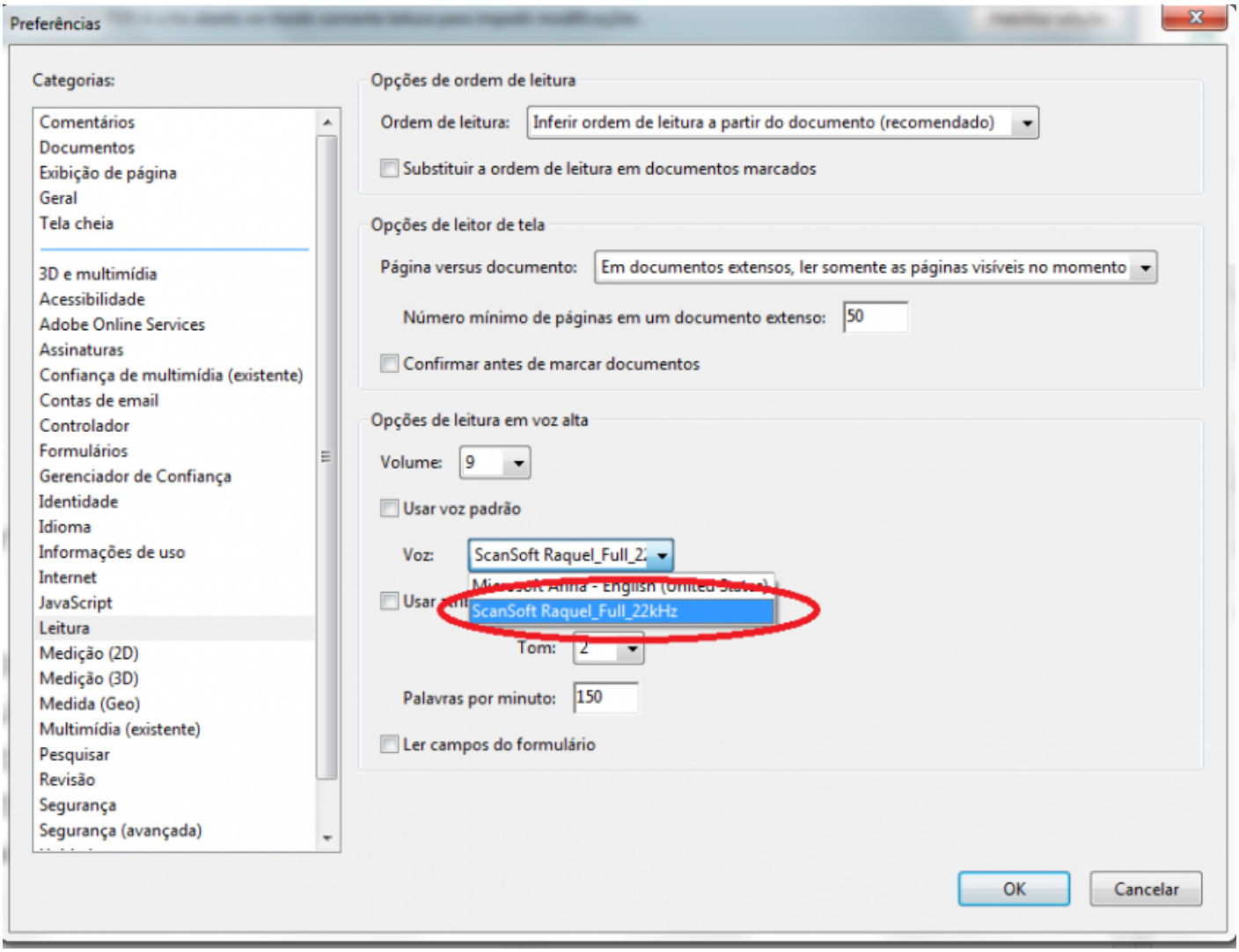
> O OCRmyPDF pode levar alguns minutos para processar o arquivo PDF.

Visualização (texto)

O arquivo com reconhecimento OCR e padrão PDF/A-1b ou 2b, aparece da seguinte forma em softwares como Adobe Reader. Como é possível notar, o texto é selecionável.

Visualização (áudio da leitura)

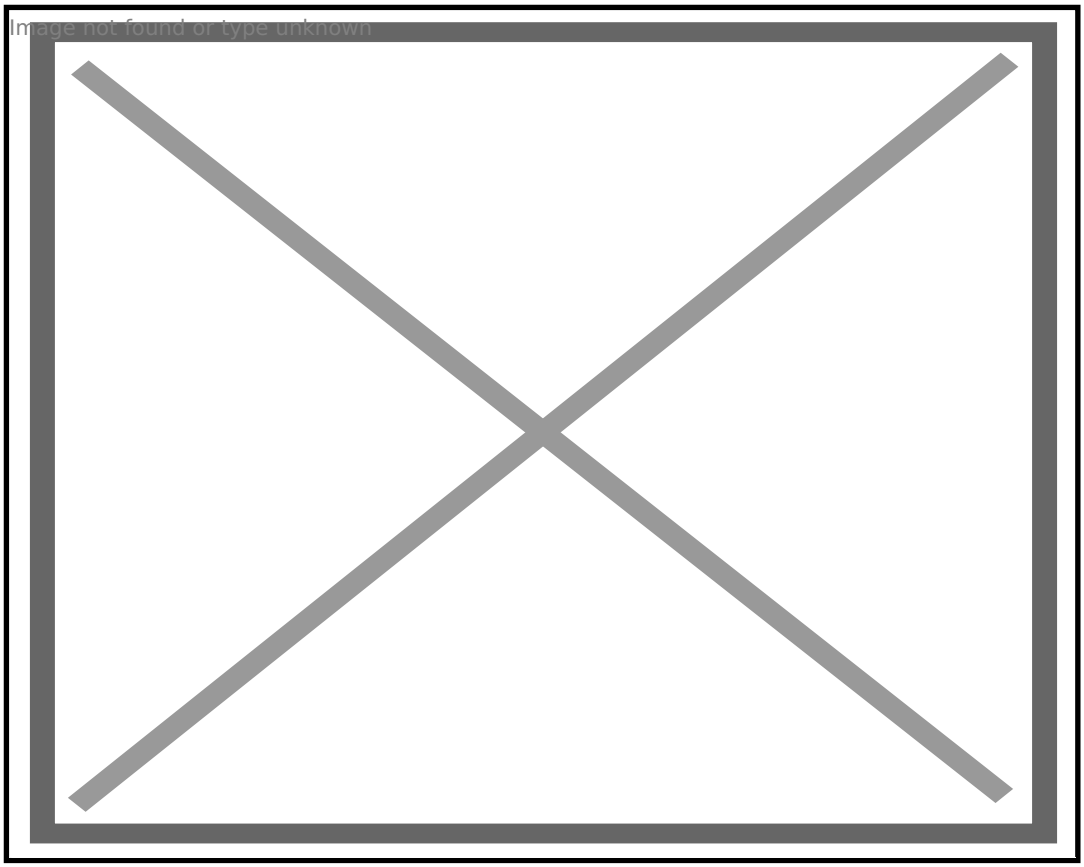
Utilizando o software Adobe Reader como exemplo, acesse as Preferências do software, ative o recurso de Leitura do texto e selecione a voz “Raquel” para uso.



Para escutar a leitura do texto pelo software, pressione os atalhos Shift + Ctrl + Y (Ativar leitura em voz alta) e Shift + Ctrl + B (Ler todo o documento). Também, é possível escutar clicando no texto ou percorrendo com o cursor do mouse.

Leitura do arquivo com OCR em ação:

Image not found or type unknown



https://www.youtube.com/watch?v=xrBu4fhoXFE&feature=youtu.be

7. Referências

https://pt.wikipedia.org/wiki/PDF/A

https://softwarepublico.gov.br/archives/thread/sei-negocio/como-criar-pdfa-conforme-iso-19005-12005

http://www.admin-magazine.com/Archive/2015/26/Look-for-file-changes-and-kick-off-actions-with-Watchman/(offset)/3

https://ocrmypdf.readthedocs.io/en/latest/index.html

http://facebook.github.io/watchman/

https://stackoverflow.com/questions/1659147/how-to-use-ghostscript-to-convert-pdf-to-pdf-a-or-pdf-x

https://www.pdf-online.com/osa/validate.aspx

http://www.tecassistiva.com.br/servicos?id=105

Revisão #48
Criado 30 November 2017 19:40:53 por Aleciano Ferreira
Atualizado 20 August 2020 00:55:35 por Aleciano Ferreira