D'après l'article de **MB** sur https://www.mathematex.fr/viewtopic.php?f=6&t=17649

Manipulation de fichiers pdf en ligne de commande

Ce sujet regroupe un certain nombre de commandes permettant la manipulation de fichiers pdf depuis un terminal linux.

1- Obtenir des informations sur un fichier pdf.

2- Apporter des modifications à un fichier pdf.

3- Fusionner des fichiers pdf.

4- Scinder un fichier pdf.

5- Extraire les images d'un fichier pdf ou en insérer dans un fichier pdf.

6- Transformer un fichier pdf en fichier texte ou en image.

7- Rechercher du texte dans des fichiers pdf.

1- Obtenir des informations sur un fichier pdf

- Afficher les métadonnées du fichier `src.pdf` avec`exiftool`...

```
$ exiftool src.pdf
```

- Afficher des informations générales sur le fichier `src.pdf` avec `pdfinfo`... On retrouve certaines informations présentes dans les métadonnées, mais on obtient quelques informations supplémentaires, dont le format de page (Page size). La seconde commande permet d'obtenir des informations spécifiques sur le format des pages 1 à 3.

```
$ pdfinfo src.pdf
$ pdfinfo -f 1 -l 3 src.pdf
```

- Lister toutes les fontes utilisées par le fichier `src.pdf` avec `pdffonts` ... On pourra par exemple identifier les fontes embarquées dans le fichier pdf (`emb=yes`). Il est également possible d'utiliser les options `-f` et `-l` pour limiter les pages à examiner.

```
$ pdffonts src.pdf
```

- Lister toutes les images contenues dans le fichier `src.pdf` avec `pdfimages` ... Il est également possible d'utiliser les options `-f` et `-l` pour limiter les pages à examiner.

```
$ pdfimages -list src.pdf
```

2- Apporter des modifications à un fichier pdf

- Modifier une métadonnée du fichier `src.pdf` avec `exiftool`... La commande suivante permet par exemple de modifier le titre du fichier.

```
$ exiftool -Title="Nouveau titre" src.pdf
```

- Compresser un fichier `src.pdf` avec `ghostscript`... L'option `-dPDFSETTINGS=/screen` correspond à la compression la plus élevée, mais on peut utiliser les options `/ebook`, `/printer` ou `/prepress` qui présentent une qualité de plus en plus élevée. Voir ici pour plus de détails sur cette option.

```
$ gs -dNOPAUSE -dBATCH -sDEVICE=pdfwrite -dPDFSETTINGS=/screen -
    sOutputFile=dst.pdf src.pdf
```

- Embarquer toutes les fontes du fichier `src.pdf` avec `ghostscript`... Comme indiqué ici, on a par défaut `-dEmbedAllFonts=true`, sauf lorsque `-dPDFSETTINGS=/screen` (ce qui n'est pas le cas dans le commande suivante).

```
$ gs -dNOPAUSE -dBATCH -sDEVICE=pdfwrite -dEmbedAllFonts=true -
    sOutputFile=dst.pdf src.pdf
```

- Redimensionner le fichier `src.pdf` au format A4 avec `ghostscript`... Dans la seconde commande, l'option `-dFIXEDMEDIA` va fixer l'orientation des pages (si elles n'avaient pas déjà toutes la même).

```
$ gs -dNOPAUSE -dBATCH -sDEVICE=pdfwrite -sPAPERSIZE=a4 -dFitPage -
    sOutputFile=dst.pdf src.pdf
$ gs -dNOPAUSE -dBATCH -sDEVICE=pdfwrite -sPAPERSIZE=a4 -dFitPage -
    dFIXEDMEDIA -sOutputFile=dst.pdf src.pdf
```

- Linéariser le fichier `src.pdf` avec `ghostscript`...

```
$ gs -dNOPAUSE -dBATCH -sDEVICE=pdfwrite -dFastWebView -sOutputFile=
    dst.pdf src.pdf
```

- Linéariser le fichier `src.pdf` avec `qpdf`...

```
$ qpdf src.pdf --linearize dst.pdf
```

3- Fusionner des fichiers pdf

- Fusionner les fichiers `src-1.pdf` et `src-2.pdf` avec `ghostscript`...

```
$ gs -dNOPAUSE -dBATCH -sDEVICE=pdfwrite -sOUTPUTFILE=dst.pdf src
    -01.pdf src-02.pdf
```

- Fusionner les fichiers `src-1.pdf` et `src-2.pdf` avec `qpdf`... La première commande ne va conserver aucune métadonnée, alors que la deuxième va conserver celles du fichier `src-1.pdf` et que la troisième va conserver celles du fichier `src-2.pdf`.

```
$ qpdf --empty --pages src-1.pdf src-2.pdf -- dst.pdf
$ qpdf src-1.pdf --pages . src-2.pdf -- dst.pdf
$ qpdf src-2.pdf --pages src-1.pdf . -- dst.pdf
```

- Fusionner les fichiers `src-1.pdf` et `src-2.pdf` avec `pdfunite`... La syntaxe est très simple, mais le fichier obtenu est plus lourd que ceux obtenus via les méthodes précédentes.

```
$ pdfunite src-1.pdf src-2.pdf dst.pdf
```

4- Scinder un fichier pdf

- Extraire les pages 3 à 5 du fichier `src.pdf` avec `ghostscript`...

```
$ gs -dNOPAUSE -dBATCH -sDEVICE=pdfwrite -dFirstPage=3 -dLastPage=5
  -sOUTPUTFILE=dst.pdf src.pdf
```

- Extraire les pages 3 à 5 du fichier `src.pdf` avec `qpdf`... La première commande ne va conserver aucune métadonnée, alors que la deuxième va conserver celles du fichier `src1.pdf`.

```
$ qpdf --empty --pages src.pdf 3-5 -- dst.pdf
$ qpdf src.pdf --pages . 3-5 -- dst.pdf
```

5- Extraire les images d'un fichier pdf ou en insérer dans un fichier pdf

- Extraire toutes les images du fichier `src.pdf` avec `pdfimages`... Le nom de fichier des images extraites par la commande suivante va débuter par img et sera suivi d'un numéro.

```
$ pdfimages -all dst.pdf img
```

- Insérer les images `img-1.png` et `img-2.png` dans le fichier `dst.pdf` avec montage... L'option `-page a4` permet de faire sorte que le fichier `dst.pdf` soit au format a4, alors que l'option `-gravity center` permet de centrer le contenu de la page, que l'option `-tile 1x` indique que les images sont alinée verticalement et que l'option `-geometry 500x+0+5` permet de fixer la largeur des images à 500 pixels et d'ajouter un espace de 5 pixels en haut et en bas de chaque image.

```
$ montage img-1.png img-2.png -page a4 -gravity center -tile 1x -
  geometry 500x+0+5 dst.pdf
```

6- Transformer un fichier pdf en fichier texte ou en image

- Transformer le fichier `src.pdf` en fichier texte avec `pdftotext`... La commande suivante va créer un fichier `src.txt` et l'option `-layout` tente de préserver la disposition des éléments.

```
$ pdftotext -layout src.pdf
```

- Transformer le fichier `src.pdf` en fichiers png avec `pdftopng`... Les commandes suivantes vont convertir chaque page du fichier src.pdf en un fichier png dont le nom va débuter par img sera suivi d'un numéro. L'option `-gray` permet de générer une image en niveaux de gris.

```
$ pdftopng src.pdf img
$ pdftopng -gray src.pdf img
```

7- Rechercher du texte dans des fichiers pdf

- Rechercher une chaîne de caractères dans des fichiers pdf avec `pdfgrep`... La première va rechercher la chaîne de caractères 'chaîne à trouver' dans tous les fichiers du répertoire courant en ignorant la case (-i) et en indiquant le numéro de la page (-n). La seconde commande va également explorer les sous-dossiers du répertoire courant (-r).

```
$ pdfgrep -in 'chaîne à trouver' *.pdf
$ pdfgrep -inr 'chaîne à trouver' *.pdf
```