

**HACK**
#50**Extraire des pages d'un PDF***Extrayez rapidement des pages d'un PDF.*

Ce hack présente un exemple concret de l'utilisation du module `PDF::Extract` de Perl et consiste en un script qui vous permettra d'extraire les pages d'un fichier PDF. En pratique, vous avez peut-être rédigé un long document (thèse, rapport, livre, etc.), et vous voulez maintenant diffuser sa table des matières séparément.

Ce script en ligne de commande effectuera cette tâche en quelques secondes et en toute simplicité. Il fonctionne sous Linux et sous Mac OS X sans modifications. Il devrait également fonctionner sous Windows si vous avez installé ActivePerl (voir [Automatiser Acrobat grâce à Perl sous Windows, \[Hack #95\]](#)). Il requiert néanmoins le module `PDF::Extract`.

Vous pouvez vérifier s'il est installé sur votre système grâce à la commande suivante :

```
perl -MPDF::Extract -e 1
```

Si vous n'obtenez pas d'erreur, cela signifie que le module est bien installé sur votre système. En revanche une erreur ne signifie pas forcément que le module n'est pas présent. Peut-être est-il quelque part sur le disque dur mais, ne se trouvant pas dans le path de Perl (que vous pouvez afficher avec la commande `perl -e "print qq(@INC)"`), il ne lui est pas accessible.

Pour de plus amples information sur les vérifications à effectuer ainsi que les instructions complètes relatives à l'installation des modules, reportez-vous à la documentation `modinstall` :

```
$ perl doc perlmodinstall
```

Le code

```
#!/usr/bin/perl
use PDF::Extract;
my $pagerange = shift @ARGV;
if ($pagerange =~ m/^(h|H|-+?h|-+?H)/) {
    print "\n pdfextract:\n emploi :\n";
    print " % pdfextract \"1,2,3,5-8\" mondocument.pdf\n";
    print " NOTEZ que les numéros de page que vous indiquez sont ,\n";
    print " relatifs au document et diffèrent peut-être des numéros \n";
    print " affichés sur les différentes pages dans Acrobat (ou n'importe \n";
    print " quel autre visualiseur de PDF, tel que gv ou Aperçu.app) \n";
    exit;
}

foreach $file (@ARGV) {
    print "Extraire les pages $pagerange de $file...\n";
    my $pdf = new PDF::Extract(PDFDoc=>$file, PDFPages=>$pagerange);
    unless ($pdf->savePDFExtract) {
        die "$0: Avertissement: impossible d'extraire les pages $file\n";
    }
}
```

Extraire des pages d'un PDF

La commande suivante génère un fichier nommé *mondocument2_3_5..7.pdf* et contenant les pages en question.

```
$ pdftextract "2,3,5-7" mondocument.pdf
```

Vous pouvez extraire en une seule fois les pages de plusieurs documents. Il suffit de saisir les chemins vers les différents fichiers sources de la manière suivante :

```
$ pdftextract "1-10" ch01.pdf ch02.pdf
```

Autant de fichiers PDF sont générés que de fichiers fournis en entrée. Dans notre exemple, deux fichiers de 10 pages sont générés : le premier contient les 10 premières pages du fichier *ch01.pdf*, le second celles de *ch02.pdf*.

Les numéros de page se réfèrent au fichier PDF lui-même. Par exemple, si votre PDF contient des pages en numérotation romaines au début, la page I sera la première page du document. De même si les premières pages de votre document ne sont pas numérotées, etc.

Voir aussi

Le site de CPAN (<http://www.cpan.org/>).
