
Comment
ça marche
.net


Beta
CodeS-SourceS
www.codes-sources.com

Haut

Extraction du contenu d'un document pdf (pdf-1.4)

18 avis

Juin 2014

Suite à une question du forum, je me suis demandé comment est-ce que fonctionnait un fichier PDF...
Si on regarde la source d'un fichier PDF dans un éditeur du type notepad++, on y trouvera différentes informations. La première étant la version du fichier. Exemple: **%PDF-1.4** %Çì ¢ 5 0 obj <</Length 6 0... Plus loin, on trouve énormément de signes bizarres, il s'agit du contenu du PDF compressé et, en bas de fichier on trouvera des lignes comme: /CreationDate(D:20071230102921+01'00') /ModDate(D:20071230102923) /Title(\376\377\000P\000D\000F\000 \000d\000e\000 \000t\000e\000s\000t) ... Il n'est pas difficile de se rendre compte que nous avons là les propriétés du PDF, écrites lettre par lettre, chacune commençant par \000. Nous pouvons donc déjà récupérer les propriétés CreationDate, ModDate, Title, Creator, Author, Keywords et Subject en faisant quelques manipulations (cf code plus bas). Pour récupérer le texte, il faut d'abord l'isoler: il est entouré des balises stream et endstream: stream xoUËN[1...-~xô!OÓ>oiô>ûÑw,Rúî«endstream Une fois isolé, il faut le décompresser. Dans le cas présent la compression est standard (on trouve une balise Filter en regardant le contenu du PDF, cette dernière indique la compression utilisée) <</Length 6 0 R/Filter /FlateDecode>> (cf http://en.wikipedia.org/wiki/DEFLATE_%28algorithm%29) Après l'avoir décompressée en utilisant gzuncompress, le contenu du fichier PDF apparaît au format PS. Si on y regarde de plus près, on y voit des lignes comme celle-ci: [(A)-2.7801(u)-2.7801( )-2.77991(p)-2.77991(l)-2.77991(a)-2.77954(i)-2.77991(s)-2.78065(i)-2.77991(r)-2.77991(,)600]TJ En observant cette ligne, on remarque entre parenthèses des lettres qui, une fois isolées, font la phrase "Au plaisir". En systématisant ce principe, on récupère tout le texte du PDF! Voici un essai que j'ai fait chez moi. J'ai cherché à extraire les données du PDF suivant :

En exécutant le bout de code en fin de page, j'ai obtenu ceci: Date de création: Sunday, 30 December 2007 10:29:21
Date de modification: Sunday, 30 December 2007 10:29:23 Titre du fichier: PDF de test Création du fichier: PDFCreator Version 0.9.3 Auteur: Malik7934 Mots clés: (non renseigné) Sujet: Un petit bonjour Contenu textuel: new 2 dimanche, 30. décembre 2007 10:28 Bonjour, Voici un fichier PDF tout simple Au plaisir, Malik7934 Et voici le code:

```
<?php

function openPDF($filename){
   $handle = fopen($filename,"rb");
   $content = fread($handle, filesize($filename));
   fclose($handle);
   return $content;
}
```

```
function versionPDF($content){
   $p1 = strpos($content,'%')+1;
   $p2 = strpos($content,'%',$p1)-2;
   $version=substr($content,$p1,$p2);
   return strcmp(trim($version),'PDF-1.4');
}

function getPDFDate($type){
   if (strcmp($type,'cre')==0)    $input = '/CreationDate';
   if (strcmp($type,'mod')==0)    $input = '/ModDate';
   global $content;
   $p1 = strpos($content,$input)+1;
   $p2 = strpos($content,0x2F,$p1);
   $line = substr($content,strlen($input)+$p1,$p2-$p1-2-strlen($input));
   $row = substr($line,2,14);
   $date = date('l, d F Y
H:i:s',mktime(substr($row,8,2),substr($row,10,2),substr($row,12,2),substr($row,4,2),substr($ro
w,6,2),substr($row,0,4)));

   return $date;
}

function getPDFInfo($type){
   if (strcmp($type,'tit')==0)    $input = '/Title';
   if (strcmp($type,'cre')==0)    $input = '/Creator';
   if (strcmp($type,'aut')==0)    $input = '/Author';
   if (strcmp($type,'key')==0)    $input = '/Keywords';
   if (strcmp($type,'sub')==0)    $input = '/Subject';
   global $content;
   $p1 = strpos($content,$input)+1;
   if (strcmp($type,'sub')==0){ $s1 = '>>';$s2 = 1;}
   else{ $s1 = 0x2F; $s2 = 2;}
   $p2 = strpos($content,$s1,$p1);
   $line = substr($content,strlen($input)+$p1,$p2-$p1-$s2-strlen($input));
   $line = preg_replace("/\[0-9]{3}/",",$line);
   if (trim($line)=='') $line = '(non renseigné)';

   return $line;

}

function getPDFContent($content,$start,$end){
   $s1 = 0;
   $e1 = 0;
   $objc = array();
```

```
    while ($s1 !== false && $e1 !== false){
        $s1 = strpos($content,$start,$e1)+(strlen($start)+1);
        if ($s1 !== false){
            $e1 = strpos($content,$end,$s1);
            if ($e1 !== false)
                $objc[] = substr($content,$s1,$e1-$s1);
        }
    }

    return $objc;
}

function extractText($content){
    $obj = array();
    $ret = array();
    $mots = array();

    preg_match_all("[([^]]+)]TJ", $content, $obj);

    for ($i=0;$i<count($obj[0])-1;$i++){
        preg_match_all("((.){1})", $obj[1][$i], $mots[$i]);

        $ret[] = implode('',$mots[$i][1]);
    }

    $ret = implode(' ',$ret);

    return $ret;

}

function getPDFText($content){
    $gc = getPDFContent($content,'obj','endobj');
    $gc_size = count($gc);
    $streams = array();
    for ($i=0;$i<$gc_size;$i++){
        $sc = getPDFContent($gc[$i],'stream','endstream');
        if (!empty($sc)){
            $streams[] = $sc;
        }
    }
    if (count($streams)>0){
        return extractText(gzuncompress(trim($streams[0][0])));
    }
```

```
}
/***********************************/

// 1. Ouverture du PDF

$filename = 'test.pdf';
$content = openPDF($filename);

// 2. Vérification de la version du PDF - seule 1.4 est supportée

if (versionPDF($content)!=0)
   echo 'Cette source n'est pas été conçue/testée pour d'autres versions que la version PDF-1.4.<br />Votre fichier a la version '.$version.', désolé';
else{

   // 3. Récupération des propriétés CreationDate,  ModDate, Title, Creator, Author, Keywords et Subject

   $creDate     = getPDFDate('cre'); echo 'Date de création: '.$creDate.'<br />';
   $modDate     = getPDFDate('mod'); echo 'Date de modification: '.$modDate.'<br />';
   $title       = getPDFInfo('tit'); echo 'Titre du fichier: '.$title.'<br />';
   $creator     = getPDFInfo('cre'); echo 'Création du fichier: '.$creator.'<br />';
   $author      = getPDFInfo('aut'); echo 'Auteur: '.$author.'<br />';
   $keywords    = getPDFInfo('key'); echo 'Mots clés: '.$keywords.'<br />';
   $sujet       = getPDFInfo('sub'); echo 'Sujet: '.$sujet.'<br /><br />';

   //4. Récupération du contenu du PDF

   $contenu     = getPDFText($content); echo 'Contenu textuel:<br />'.$contenu;
}
?>
```

Remarque importante : Ne croyez pas pouvoir utiliser ce code tel quel pour par exemple parser vos fichiers PDFs lors de recherches sur votre site. Plusieurs raisons vont l'empêcher: ce code ne fonctionne que avec PDF-1.4, il ne traite pas les cas cryptés et ne gère que du texte pur. Le but de ce code est de mettre sur la voie ceux qui seraient intéressés d'aller plus loin en observant le contenu des PDFs et en lisant la doc proposée par Adobe :-) En espérant que ce petit code/tutoriel serve à quelqu'un!!! Au plaisir, Malik7934

Commentaires Ajouter un commentaire

Afficher les 18 commentaires

Ajouter un commentaire