



PDF Parser (/)

Home (/)

Documentation (/documentation)

License (/license)

Demo (/demo)

Contact (/contact)

Donate (/donate)

PDF Parser (/)

PHP library to parse PDF files and extract elements like text. (/)

build **passing** (<https://travis-ci.org/smalot/pdfparser>) downloads **1.27 M** (<https://packagist.org/packages/smalot/pdfparser>)
stable **v0.14.0** (<https://github.com/smalot/pdfparser>)

Download from GitHub
(<https://github.com/smalot/pdfparser>)

Use it from Packagist
(<https://packagist.org/packages/smalot/pdfparser>)

Home (/) / Documentation

Documentation

PdfParser, a standalone PHP library, provides various tools to extract data from a PDF file.
Currently, secured documents are not supported.

This Library is still under active development. As a result, users must expect BC breaks when using the master version.

This project is supported by Actualys (<http://www.actualys.com/>).

Prerequisites

This library requires **PHP 5.3**.
PDFParser is built on top of TCPDF parser.
This library will be automatically downloaded through Composer command line.

Installation

Using Composer

Add PDFParser to your composer.json file :

```
1. {  
2.     "require": {  
3.         "smalot/pdfparser": "*"   
4.     }  
5. }
```

Now ask for composer to download the bundle by running the command:

```
1. $ composer update smalot/pdfparser
```

As standalone library

First of all, download the library from Github by choosing a specific release (<https://github.com/smalot/pdfparser/releases>) or directly the master (<https://github.com/smalot/pdfparser/archive/master.z>).
Once done, unzip it and run the following command line using composer (<http://getcomposer.org/download/>).

```
1. $ composer update
```

This command will download any dependencies (*Atoum library*) and create the 'autoload.php' file.

Now create a new file with this content, in the same folder :

```

1. <?php
2.
3. // Include 'Composer' autoloader.
4. include 'vendor/autoload.php';
5.
6. // Your code
7. // ...
8.
9. ?>

```

Unit tests with Atoum

Run Atoum unit tests (with code coverage - if xdebug installed) :

```
1. $ vendor/bin/atoum -d vendor/smalot/pdfparser/src/SmaLot/PdfParser/Tests/
```

Once this command is ended, the folder "coverage/" will contain html pages with a code coverage summary.

Use

This sample will parse all the pdf file and extract text from each page.

```

1. <?php
2.
3. // Include Composer autoloader if not already done.
4. include 'vendor/autoload.php';
5.
6. // Parse pdf file and build necessary objects.
7. $parser = new \SmaLot\PdfParser\Parser();
8. $pdf    = $parser->parseFile('document.pdf');
9.
10. $text = $pdf->getText();
11. echo $text;
12.
13. ?>

```

You can too extract text from each page handy or for a specific page.

```

1. <?php
2.
3. // Include Composer autoloader if not already done.
4. include 'vendor/autoload.php';
5.
6. // Parse pdf file and build necessary objects.
7. $parser = new \SmaLot\PdfParser\Parser();
8. $pdf    = $parser->parseFile('document.pdf');
9.
10. // Retrieve all pages from the pdf file.
11. $pages = $pdf->getPages();
12.
13. // Loop over each page to extract text.
14. foreach ($pages as $page) {
15.     echo $page->getText();
16. }
17.
18. ?>

```

Here a sample code to extract metadata from document (Author, Creator, CreationDate, ...).

```
1. <?php
2.
3. // Include Composer autoloader if not already done.
4. include 'vendor/autoload.php';
5.
6. // Parse pdf file and build necessary objects.
7. $parser = new \Smalot\PdfParser\Parser();
8. $pdf = $parser->parseFile('document.pdf');
9.
10. // Retrieve all details from the pdf file.
1. $details = $pdf->getDetails();
2.
3. // Loop over each property to extract values (string or array).
4. foreach ($details as $property => $value) {
5.     if (is_array($value)) {
6.         $value = implode(' ', $value);
7.     }
8.     echo $property . ' => ' . $value . "\n";
9. }
10.
1. ?>
```

You can contact and follow me on Twitter @sebastienmalot (<https://twitter.com/sebastienmalot>)
Code licensed under GPLv2 - This project is supported by Actualys (<http://www.actualys.com/>).

GitHub (<https://github.com/smalot/pdfparser>) • free(code) (<https://freecode.com/projects/pdfparser>) • Packagist (<https://packagist.org/packages/smalot/pdfparser>) • Travis CI (<https://travis-ci.org/smalot/pdfparser>)