



**Workshop: PDF Treatment**

**Prepared by: TheCodeHolics**

Table of Contents

[What is a pdf ? 3](#_Toc153754273)

[Why pdf ? 4](#_Toc153754274)

[Tasks 5](#_Toc153754275)

[Extracting data from pdf 5](#_Toc153754276)

[Pdf metadata removing 6](#_Toc153754277)

[Code snippets 6](#_Toc153754278)

[Converting Pdf to anther formats 8](#_Toc153754279)

[Pdf to audio 9](#_Toc153754280)

[Demo 11](#_Toc153754281)

# What is a pdf ?

PDF is an abbreviation that stands for Portable Document Format. It's a versatile file format created by Adobe that gives people an easy, reliable way to present and exchange documents - regardless of the software, hardware, or operating systems being used by anyone who views the document.

The PDF format is now an open standard, maintained by the International Organization for Standardization (ISO). PDF docs can contain links and buttons, form fields, audio, video, and business logic. They can be signed electronically and can easily be viewed on any Operating system.

# Why pdf ?

The use of PDF offers several advantages, including:

1. **Portability**: PDF files can be read anywhere and on any device, making them highly accessible

2. **Compactness**: They are smaller in size, take up less space, and open faster, making it easier to share and transfer large files

3. **Security**: PDFs can be encrypted and offer secure ways to send personal or confidential data.

4. **Universal format**: PDFs can be viewed on most devices and are widely adopted globally.

5. **Graphic integrity**: They preserve the layout and content across all devices, ensuring the visual appeal of the documents.

6. **Searchability**: Users can easily find desired content through quick searches within PDF documents.

7. **Interactivity**: PDFs can feature hyperlinks, file attachments, and multimedia content, making them versatile for various types of content.

These benefits make PDFs an ideal choice for sharing, collaborating, and preserving the integrity and security of documents.

# Tasks

## Extracting data from pdf

Extracting data from a PDF involves retrieving specific content like text and images from the file. This process separates and retrieves textual information and graphical content such as pictures or diagrams embedded within the PDF document.

* **Steps to Extract Images and Text from a PDF**

First of all, download Git & Composer:

So, you can use Git and Composer to download smalot/pdfparser using this command in the terminal:



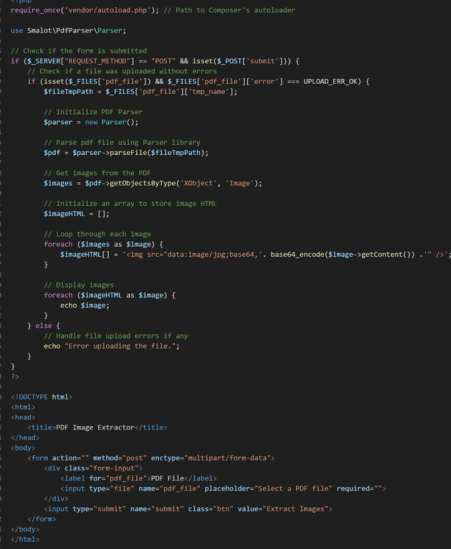
Finally, write a PHP script using smalot/pdfparser to extract images and texts from a PDF.

Code snippets:

Extracting text:



Extracting Picture (jpg):



## Pdf metadata removing

What it does: uploading PDF files and modify or remove metadata

Code snippets

* library import :
* load pdf file and display its metadata on the textbox fields in html page :
* modify of metadata and re-download the pdf file :

## Converting Pdf to anther formats

## Pdf to audio

Requirements : **NodeJS**

In the context of the task, the process involves several steps:

1. Input: The expected input is a PDF document.
2. Text Extraction: The text will be extracted from the PDF.



1. Language Detection: The language of the extracted text will be detected.

4. Text-to-Speech Conversion: Finally, a request will be made using the Google Text-to-Speech API to convert the extracted text into audio.



# Demo