

## Join GitHub today

Dismiss

GitHub is home to over 28 million developers working together to host and review code, manage projects, and build software together.

[Sign up](#)

Library focussing on reliable PDF Parsing

157 commits

2 branches










0 releases

2 contributors

Branch: master ▾

[New pull request](#)[Find file](#)[Clone or download ▾](#)

 jdehaan Merge pull request #12 from carbon/master ... Latest commit 53bac43 19 hours ago

 .vscode	Fixes #7 development settings for vscode	4 days ago
 PdfInfoTool	Remove private obj files from repository	a day ago
 SafeRapidPdf.UnitTests	Rename Adapter.cs to StringExtensions.cs	a day ago
 SafeRapidPdf	Format PdfObject	a day ago
 .gitignore	Ignore .vs	3 days ago
 .travis.yml	#9 well then stay at .NET core 2.0.0	4 days ago
 LICENSE.md	#6 could not yet manage to port, keeping license text	5 days ago
 README.md	Fixes a space slip...	3 days ago
 SafeRapidPdf.sln	Added PdfInfoTool to test the library in batch scripts and bigger pdfs	12 days ago
 ca.ruleset	Code cleanup	5 days ago

 README.md

# SafeRapidPdf

## CI-Status

master passing

## Introduction

There is already a very good pdf parser and generator: [itextsharp](#). But it doesn't focus on parsing and its licensing model makes it inappropriate for some purposes. This designed and developped from scratch library is provided under the liberal MIT license (Refer to details in the License section).

The focus of the library is on reading and parsing, not on writing.

The goals followed are:

- parsing and analysing PDF contents (virus check for example)
- integrity of parsing (document scans from start to end gathering all objects)

- no quirks, invalid PDFs are not parsed
- allow extraction of text and images at a very low level

This library is not intended for following purposes:

- rendering a PDF
- modifying a PDF
- generating a PDF

## File structure

---

This library attempts to provide a quick and yet reliable parser for PDF files. It focusses on an integral parsing of the whole PDF into its primitive objects.

- Strings
- Numeric values
- Booleans
- Streams
- Arrays
- Dictionaries
- Indirect Objects
- Indirect References
- Cross Reference sections

## Document structure

---

The interpretation layer allows then a decomposition into pages and images among other high level objects.

- Cross reference table
- Root
- Pages
- Graphics
- Text
- Fonts

The library is not interested in rendering the PDF only the informative parts will be extracted such as the position and size of text and graphics for example.

## Online resources

---

- Wikipedia explanations on [the PDF format](#)
- A python library with similar goals: [pdf-parser](#)

It is recommended to read the specification of the PDF language 1.7 for a deeper insight.

## Authors

---

The SafeRapidPdf contributors:

- Jaap de Haan (initiator)

## License

---

The MIT license (Refer to the [LICENSE.md](#) file)

