

Considering PyMuPDF

Evaluation of PDF Rendering Engine Change November 2023



Overview

- Code impact
- Installation impact
- Metrics: conversion speed, size and success rate
- Rendering Impact (manual analysis)
- License Impact



Code Impact

- Simplifies significantly server code
- Replaces multiple tools:
 - Convert (GraphicsMagick)
 - PDFtoPPM (poppler-tools)
 - ps2pdf (GhostScript for compression)
 - pdfunite (for merging PDFs)
- Integration with Tesseract-OCR
 - Whatever it does it's faster than our use of tesseract
- No more need for subprocess calls except LibreOffice

dangerzone/conversion/common.py	+7 - 0
dangerzone/conversion/doc_to_pixels.py	+27 -125
 dangerzone/conversion/pixels_to_pdf.py 	+43 -106

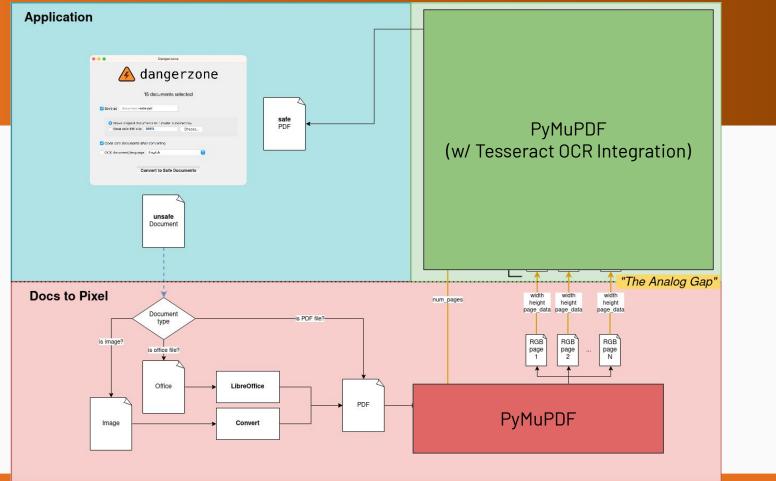


Code Impact

[big] No bottlenecks: no need to write files to disk (like PDFtoPPM) in first conversion

This is what made us consider PyMuPDF in the first place (see <u>#616</u>)





A Dangerzone

https://dangerzone.rocks

Docker Build Challenges

- Fetches <u>MuPDF binary</u> from MuPDF.com
- ! Does not verify it
- Increases container build time significantly (though we may be able to reduce it)



Metrics: Performance

Faster on larger files, slower on smaller files

	Test Data			Results			
Branch	Commit	Test Type	Test Device	Command	average	Performance boost (versus main)	Failure Rate
53115b3	443-stream-pa	214 page (no OCR)	Dell XPS 13 (9305)	time python ./dev	45.5502	1.36	
04f8fa8	main	214 page (no OCR)	Dell XPS 13 (9305)	time python ./dev	62.111	1.00	5
53115b4	443-stream-pa	214 page (OCR)	Dell XPS 13 (9305)	time python ./dev	706.8675	1.41	5 I
04f8fa9	main	214 page (OCR)	Dell XPS 13 (9305)	time python ./dev	1201.3112	1.00	-
04f8fa83	443-stream-pa	214 page (no OCR)	M1 Mac	time python ./dev	45.26366667	1.17	-
6876fa4	main	214 page (no OCR)	M1 Mac	time python ./dev	54.2516	1.00	2
04f8fa83	443-stream-pa	214 page (OCR)	M1 Mac	time python ./dev	228.0674	1.33	2
6876fa5	main	214 page (OCR)	M1 Mac	time python ./dev	340.2052	1.00	-
53115b4	443-stream-pa	4000 docs test (no O	M1 Mac	make test-large	9740.88	0.79	1.33%
04f8fa9	main	4000 docs test (no O	M1 Mac	make test-large	8080.741	1.00	0.77%



https://dangerzone.rocks

Metrics: Document Size

Shrunk: 4762, average: -34.4%

Increased: 133, average: 68.5%

Total average: -4.3%



https://dangerzone.rocks

Bonus points: Less Compression Artifacts

Note: low resolution in PyMuPDF will be resolved by <u>#626</u> original (adobe acrobat)

The background should be gray.

PDFtoPPM (ps2pdf) Compression artifacts The background should be gray. Yes, also font differences

The background should be gray.

Other PyMuPDF honorable mentions

- **Supports more file formats** (XPS, EPUB, MOBI, FB2, CBZ, SVG and various image formats), which mean we can add these at any time with no extra cost
- Opens possibility of on-host pixels2pdf

Dropping extra binary dependencies (poppler, ps2pdf, pdfunite) opens up path for on-host pixels2pdf. We only need to figure out Tesseract-OCR install on Windows and MacOS and PyMuPDF in the old ubuntu 20.04



Open Questions

- Container image size impact:
 - Does MuPDF already include CJK so we can remove its dependencies
- Can our vuln. Scanner (grype) detect PyMuPDF?



WARMUP

PDF is technically a standard but rendering can vary widely

Akin to the web standards

WARMUP

PDF is a standard but...

This particular issue seems to be the lack of font embedding in the

document

PDFtoPPM ' (6&5,37,21 **PyMuPDF** F GUE T KRVKQ P Adobe Acrobat S Cannot find or create the font 'CenturyGothic,Bold'. Some characters may not display or print correctly.

Evince ' (6&5,37,21

GitHub.com rendering DESCRIPTION

MacOS Preview

DESCRIPTION



PDFs Visual Diffing

Using the tool github.com/ vslavik/diff-pdf

"AaBbCc"

Main Branch Using pdftoppm for PDF rendering

PyMuPDF branch Uses PyMuPDF for PDF rendering

"AaBbCc"

"AaBbCc"

Methodology
1. Run large test on <10KB docs from this set and save results
2. Repeat on the other branch for comparison
3. Manual inspection of diffs (data here)

File Name main (pdftoppm) pymupdf2 branch			h
./simple3.pdf-safe.pdf	64331	29108	widely different
./issue4722.pdf-safe.pdf	9716	6697	text widely differ:
./tdf135035.odt-safe.pdf	18620	13143	some lines are fi
./tb-rl-textbox.odg-safe.pdf	45911	31511	rotation
./tbrl-frame.odt-safe.pdf	17389	11676	rotation
./tdf108482.odt-safe.pdf	22383	28316	rotation
./tdf130657.xlsx-safe.pdf	59897	43520	rotation
./tdf136364.xlsx-safe.pdf	24264	17423	rotation

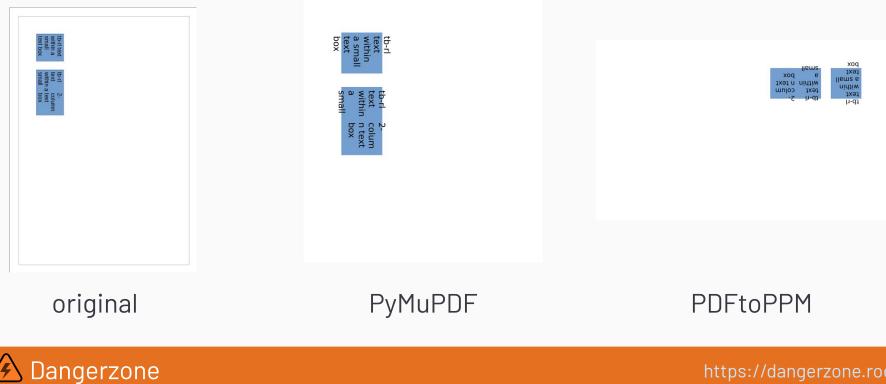
Rotations

There were lots of them in many different file types.

PyMuPDF seems to have always the rotation right.

(かたんだせべつの音楽にいなのなった)の、 能力を含むして、 ないないのなった。 物の協議能にないな部業のなくないたいの、 ないないないで、 ないないないで、 そうないないないないないない。 認識力だれたのの知道ので、 たいないないで、 そうないないないで、 そうないないないない。 に、 たいないないないないで、 そうないないで、 そうないないないで、 そうないないないないない。 に、 たいないないないないないで、 そうないないで、 そうないないないで、 そうないないないないないないないないないない。 たいないたいないないないないないで、 そうないないで、 そうないないないないないないないないないないないないないないないない。 たいないたいないないないないで、 そうないないないで、 そうないないないないないないないないないないないないない。 たいないたいないないないないないで、 そうないないないないないないないないないないないないないないないないないないない	を の の に の に の に の の の の の の の の の の の の	日本のとほりの心象スケムの認知のという。そのとほうのものとなったもち、そのかけとなかたする方角からしたまでたもちっぷけたもち、そのかげとひかりのひとくにの心象スケムのとほうの心象スケムのとほうの心象スケムのとほうの心象スケムのとほうの心象スケムのとほうの心象スケムのとほうの心象スケムのとなったものです。	いかにもたしかにとも
(かたんだせたくつの目の少くなため他のやいに、 客型を含意の日大な気障か めの感染剤を見たすべきのでは利用した。 名のの一大な気障か。 私のなどの日本ないなどの 間違いたたかの引用していないない。 そうないは、それなど、それなどなど、それなどないなどのでは、 たための事業にいなののついの回動です。 たいないないで、 たいななどのなどので、 たための事業にいなののついの回動です。 たいななどのなどので、 たいななどのなどの たための事業にいなののついの回動です。 たいななどのなどので、 たいななどのなどの たための事業にいなののついの回動です。 たいななどのなどで、 たいななどのなどの たための事業にいなののついの回動です。 たいななどのなどで、 たいななどのなどの たための事業にいなののついの回動です。 たいななどのなどで、 たいななどのなどの たための事業にいなののついの回動です。 たいななどのなどで、 たいななどのなどの たための事業にいなののついの回動です。 たいななどのなどで、 たいななどのなどの たためのからいのの回動です。 たいななどのなどで、 たいななどのなどで、 たいななどのなどの たためのかないなどんがでした。 たいない、 たいないなどで、 たいなんでのなどの たいないたいなどのなどの。 たいない、 たい、 たい、 たい、 たい、 たい、 たい、 たい、 たい、 たい、 た	のかのたい服用です。 のに、のために、 のに、 のに、 のに、 のに、 のに、 のに、 のに、 の	本れを受らないとして感情をついない。 ・ため受けない。 をではなくもその的なななない。 なったからのではない。 なったからのではない。 したではなくられていた。 したではなくられていた。 したではないない。 したではないないです。 したではないないない。 したではないないない。 したではないないない。 したではないないない。 したではないないない。 したではないないない。 したではないないない。 したではないないない。 したではないないない。 したではないないない。 したではないないない。 したではないないないない。 したではないないないない。 したではないないないできた。 したではないないないできた。 したではないないないないできた。 したではないないないない。 したではないないないない。 したではないないないないないないない。 したではないないないないないないないないない。 したではないないないないないないないないない。 したではないないないないないないないないないないないないないないない。 したではないないないないないないないないないないないないないないないないないないない	かにともりつづけるしく明滅しながら
(かたんかせんへの色色の人々ためのやって) 物の麻癬にしたみが加いていた。 物のな薬をしたした。 物のな薬をしたした。 たただの事業にしたのとした。 たたかの事業にしたのとした。 たたかの事業にしたのとした。 たたかの事業にしたのとした。 たたかの事業にしたのとした。 たたかの事業にしたのとした。 たたかの事業にしたのとのです。 たたかの事業にしたのとした。 たたかの事業にしたのとした。 たたかの事業にしたのというでになった。 になっていたがた業を体存す。 たたかの事業にしたのとした。 たたかの事業にしたのとした。 たたかの事業にしたのとのです。 たたかの事業にしたのとした。 たたかの事業にしたのとのです。 たたかの事業にしたのとした。 たたかの事業にしたのとした。 たたかの事業にしたのとした。 たたかの事業にしたのとした。 たたかのからのため、 たたかからないた。 たたかのからのため、 たたかのからのため、 たたかからないた。 たたかからないた。 たたかからないた。 たたかからないた。 たたかからないた。 たたかからないた。 たたかからないた。 たたかからないた。 たたかからないた。 たたかからないた。 たたかからないた。 たたかからないた。 たたかからないた。 たたかからないた。 たたかからないた。 たたかからないた。 たたかからないた。 たたかからないた。 たかかかたのです。 たかかかからのです。 たかかかからないた。 たかかかからのです。 たかかからのです。 たかかかからのです。 たかかかからのです。 たかかかからのです。 たかかからのです。 たかかかからのです。 たかかかからのです。 たかかかかかかかか。 たかかかからのです。 たかかかからのです。 たかかかからのです。 たかかかからのです。 たかかかからのです。 たかかかからのです。 たかかかからのです。 たかかかからのです。 たかかかからのです。 たかかかからのです。 たかかかからのです。 たかかかかからのです。 たかかかからのでのででのでかかかかっかかかからのです。 たかかかからのでのでのでかかかかからのです。 たかかかかかかっかかからのでのでのでのででででかかかかかっかかかかかかかかかかかか			
(かたんだせたくつき目の少くなたあめのやいに 物の試解したほかくない工能(マイン・オート) かただいが出たいためのこうした」の「「「「「」」」」」」」」」」」」」」」」」」」」」」」」」」」」」」」			
	ある程度まではみんなに共通いたします それが虚無ならば虚無自身がこのとほりで 記録されたそのとほりのごのけしきで たきわしかに記録されたこれらのけしきは それらも畢竟こうちののひつの風物です それれ新鮮な本体論もかんがくませうが 一番をを決くませらが	送明な人類の日大な足跡を あるいは日間記砂岩の際面に すてきな化でを発掘したり からかかな水準素のあたりから 声音らやかな水準素のあたりから 声をすっぱいの無色な孔袋が肥たとおもひ 相思した細胞をまた大気を満まりを選出した 品質をかない。1年まぐらは開いた。	るひは修羅の十億年一點にも均しい明暗
		い 寒や時間 それ 自身の 生質 としてすべてこれ らの 命題は	
心象や時間それ自身の性質としてすべき。		第四次延長のなかで主張されます。小倉や町間でお月日にの作賞として	

Rotations example



Font Differs

There were lots of them in many different file types.

PyMuPDF seems to have always the rotation right compared to the original





Font Differences

Номер полиса

Only certain fonts differ 1 Helvetica Helvetica 2 Arial Arial Arial Arial Aria Only certain fonts differ MaBbCc MabbCc



https://dangerzone.rocks

One case of PyMuPDF weirdness

This must be a particularly problematic PDF (simple3.pdf)

Even adobe reader fails to render it!

original (apple preview)



PyMuPDF characters missing

tdf148706.odt Issue5244.pdf operator-in-TJ-array.pdf

1001 04	
1021.04	



Grandes Clientèles, Financements et Marchés

PDFtoPPM characters missing

One single case detected:

issue5954.pdf

Issue 5954

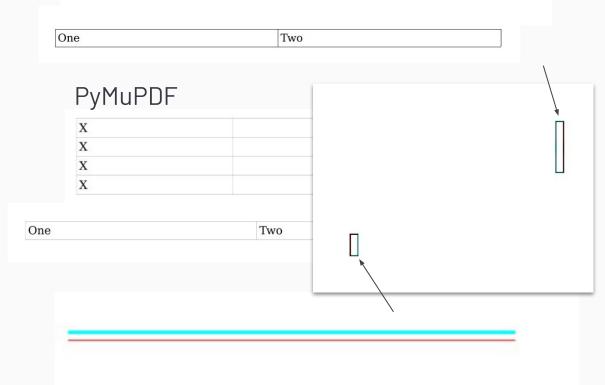
Line Rendering differs

From .ods/.odt file

veraPDF test suite 6-3-3-t01-fail-j.pdf PyMuPDF closer to Preview (adobe doesn't even render)

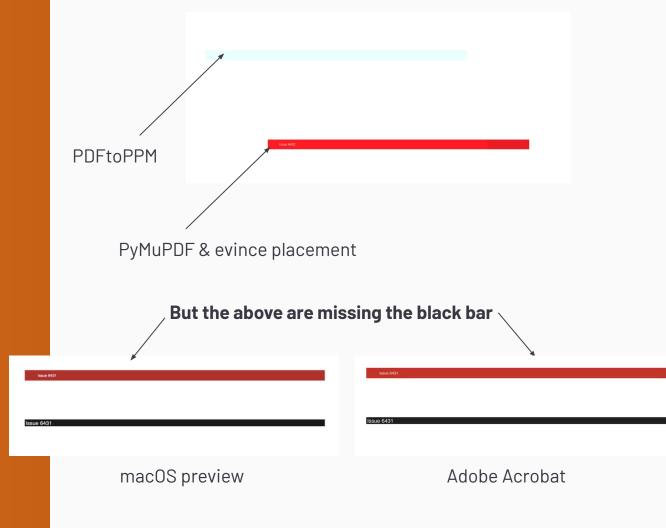
PDFtoPPM

X		
X		
X		
X		



Placement

PyMuPDF had better placement



AGPL License Impact

