Ghostscript and MuPDF Status OpenPrinting Summit May 2022

Michael Vrhel, Ph.D.

Artifex Software Inc.

Novato CA

Outline

Ghostscript

MuPDF

Ghostscript – Changes since last meeting

MuPDF – Changes since last meeting

PyMuPDF, MuJS, Web Assembly

LittleCMS2MT, GS API language bindings

Code security

Current and future work

Ghostscript

Ghostscript is a document conversion and rendering engine.

Ghostscript, GhostPCL, GhostXPS, GhostPDF, GhostPDL

Converts between and renders PDF, PS, PCL, PCL-XL, XPS

Dual license GNU AGPLv3 / Commercial

Source and documentation available at www.ghostscript.com

MuPDF

Open-source software framework for viewing and converting PDF, XPS, and e-book documents

Dual license GNU AGPLv3 / Commercial

Written in C, but has JNI bindings that work on both Oracle's Java and Android

Source and documentation available at www.mupdf.com

MuPDF

Command Line Tools

mutool draw

Primarily used for rendering a document to image files.

mutool convert

For converting documents into other formats.

mutool trace

Debugging tool used for printing a trace of the graphics device calls on a page.

MuPDF

Command Line Tools

mutool show

A tool for displaying the internal objects in a PDF file.

mutool extract

Extract images and embedded font resources.

mutool clean

Rewrite PDF file. Used to fix broken files, or to make a PDF file human editable.

MuPDF

Command Line Tools

mutool merge

Merge pages from multiple input files into a new PDF.

mutool create

Create a new PDF file from a text file with graphics commands.

mutool run

A tool for running Javascript programs with access to the MuPDF library functions.

Artifex

MuPDF

JavaScript

Examples in docs/examples:

pdf-merge.js

pdf-portfolio.js

pdf-create.js

and more…

Example:

mutool run pdf-merge.js output.pdf input1.pdf input2.pdf …

Artifex

Changes to GS since last meeting

Release 9.55 October 2021

- Includes new PDF interpreter. Invoked with –dNEWPDF=true

- Includes stand-alone PDF only binary

- pdfwrite device supports passthrough for JPX/JPG2000

- Improvements in demo apps for C, C#, Java, and Python

- Object specific halftoning

Artifex

Changes to GS since last meeting

Release 9.56 March 2022

- New PDF interpreter enabled by default (–dNEWPDF=false to disable)

- Calling Ghostscript via GS API now threadsafe. Limitation is the
  X11 devices which cannot be made threadsafe due to interaction
  with X11 server.

- PSD (Photoshop) output format includes ICC profile

- API demo added for MATLAB

- GS able to generate all formats for driverless printing.
  Apple Raster/URF cups/pwgraster pclm/pclm8 pdfwrite/pdfimage8/pdfimage24/pdfimage32

Artifex

Changes to MuPDF since last meeting

Release 1.19 October 5th 2021

Microsoft Word DOCX output format
ODT output format (OpenOffice)

Undo and redo functions
Autosave and restore functions
Reflow document wrapper

Enhancements to OCR functions
Additional 'mutool run' and Java library functions
Added new Unicode scripts and fonts for EPUB
Support opacity for all markup annotation types

C++ and Python bindings now run on Windows
Python bindings are now available as pre-built binaries
Added demo PDF viewer using the Python MuPDF bindings and PyQt

Artifex
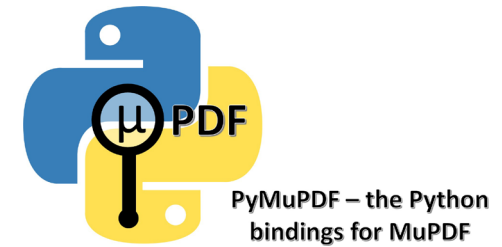
PyMuPDF

Python binding for MuPDF

PyMuPDF can be installed from Python wheels

Windows (32bit and 64bit), Linux (64bit, Intel and ARM) and Mac OSX (64bit, Intel), for Python versions 3.7 and later.

See https://pymupdf.readthedocs.io/en/latest/intro.html for detailed documentation

Heavily used in the Python community (downloaded more than 19 million times)

Dual license GNU AGPLv3 / Commercial

Little CMS2MT

We continue to use a fork of Little CMS2 that is thread safe.

Fork is currently available with git checkout of Ghostscript.

We bring in any bug fixes applied to Little CMS2 .

Developed SSE4.2, AVX2, NEON plug-in for tetrahedral interpolation with cmyk, rgb, or gray output. (Commercial license only)

MuJS

MuJS is a library, written in C.
MuJS has no notion of a main program:
it only works embedded in a host client program.
The host program can invoke functions to execute
Javascript code, read and write Javascript variables,
and register C functions to be called by Javascript.

Implements EMCAScript ECMA-262

Open source under ISC license
https://opensource.org/licenses/ISC

http://git.ghostscript.com/?p=mujs.git;a=summary

MuPDF WebAssembly

Binary instruction format for a stack-based virtual machine.

Wasm is designed as a portable target for compilation enabling deployment on the web for client and server applications.

```
mupdf.openDocument(filename)   Open a document and return a handle.
mupdf.freeDocument(doc)        Free a document and its associated resources.
mupdf.documentTitle(doc)       Return the document title as a string.
mupdf.documentOutline(doc)     Return element containing the table of contents formatted as an unordered HTML list with links to pages
mupdf.countPages(doc)          Return the number of pages in the document.
mupdf.pageWidth(doc, page, dpi)        Return the width of a page.
mupdf.pageHeight(doc, page, dpi)       Return the height of a page.
mupdf.drawPageAsPNG(doc, page, dpi)    Render the page and return a PNG image formatted as a data URI.
mupdf.pageLinks(doc, page, dpi)        Retrieve an HTML string describing the links on a page.
mupdf.drawPageAsSVG(doc, page)         Return a string with the contents of the page in SVG format.
mupdf.drawPageAsHTML(doc, page)        Return a string with the contents of the page in HTML format, using absolute positioned elements.
```

Artifex

Ghostscript API Bindings
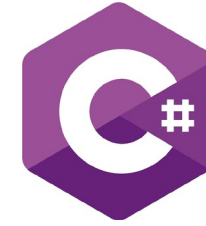
In ghostpdl/demos folder:
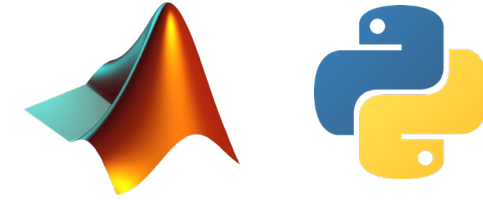
      C – Contains a VS project that exercises API

      C# – Contains simple demo viewer for Windows (WPF UI) and for Linux (MONO
              with GTK UI).   Mimics C API but has helper methods to extend API

      Java – Contains simple Java demo viewer.
            Mimics C API but has helper methods to extend API

      Python – Mimics C API.   Has demo example usage (ghostpdl\demos\python\examples.py)

      MATLAB – Uses MEX file to interface to GS API.
             Allows MATLAB function calls like:

```
page_image = gs_displaydevice(input_file, PlanarRGB, page_number, resolution);
```

Artifex

Code Security/Analysis Methods

Fuzzing of test files used to detect simple faults

Coverity : https://scan.coverity.com/projects/ghostpdl (dereferences of NULL pointers, use of uninitialized data,

memory corruptions, buffer overruns, control flow issues, incorrect expressions, unsafe signed values)

Coverage tests run periodically: https://ghostscript.com/coverage/

Various compilers used and warning report provided with every commit (gcc, clang)

Address Sanitizer: Testing for buffer overflows, dangling pointer overflows

Valgrind: Testing for buffer overflows, use of uninitialized memory/variables

Memento: Memory leak/corruption analyzer and "Memory Squeezing".   Part of Ghostscript build memento.h/c

Current/Future Work

-dBlackVector –dBlackText –dBlackThresholdL –dBlackThresholdC

These options allow the forcing of vector and text content to black (or white) even to pdfwrite device.

Improved page control for PDF files like what is available for XPS files. E.g.

    -sPageList=1,2,1,2
    -sPageList=10-5
    -sPageList=1-,-1

Preservation of PDF accessibility content for pdfwrite device

More Information

Repositories located at
git://git.ghostscript.com

Ghostscript discussions on Discord channel #ghostscript
MuPDF discussions on Discord channel #mupdf

Bug reports
bugs.ghostscript.com

Additional information at www.mupdf.com   www.ghostscript.com