Ghostscript and MuPDF Status
OpenPrinting Summit May 2021

Michael Vrhel, Ph.D.

Artifex Software Inc.

Novato CA

Outline

Ghostscript

MuPDF

Ghostscript – Changes since last meeting

MuPDF – Changes since last meeting

A few details

Current and future work

Ghostscript

Ghostscript is a document conversion and rendering engine.

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

Designed toward mobile environment use

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.

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 …
```

Changes to GS since last meeting

Release 9.52 March 2020

- Implementation of fillstroke device method (fixes issues with transparency in those operations)

- Zero Coverity issues achieved (will be maintained moving forward)

- Introduction of SIMD acceleration methods for commercial licenses (Transparency blending, interpolation, halftoning, color conversion)

- Moved to Visual Studio 2019 solution

Changes to GS since last meeting

Release 9.53 September 2020

- Bindings provided for Python, Java, and C# . Example usage is given in the demos folder – more about this later.
- Build with Tesseract OCR engine. Devices pdfocr8/pdfocr24/pdfocr32 render to an image, OCR that image, and output the image "wrapped" up as a PDF file, with the OCR generated text information included as "invisible" text.
- Allows creation of fully redacted PDF output.
- Updates made to the API. See https://www.ghostscript.com/doc/API.htm

Changes to GS since last meeting

Release 9.53 September 2020 Continued

- It was announced that OpenPrinting Vector/Raster Printer Drivers (that is, the opvp and oprp devices) deprecated and will be removed in the future.

  That decision was reversed after finding that the opvp device was being used by some venders.

- Reintroduction of the patch level to the version number. Helps facilitate handling security related issues.

Changes to GS since last meeting

Release 9.54 March 2021

- Overprint (and spot color) simulation available to all output devices, allowing quality previewing/proofing of PDF jobs that rely on overprint.

- Introduction of "docxwrite" device. Adds the ability to output to Microsoft Word "docx" format.

- The pdfwrite device can make use of Tesseract OCR engine to improve searchability and copy and paste functionality when the input lacks the metadata for that purpose.

Changes to GS since last meeting

Release 9.54 March 2021 continued

- Introduction of a "map text to black" function, where text drawn by an input job can be forced to draw in solid black.
- N-up imposition.   -sNupControl=number1xnumber2

Changes to MuPDF since last meeting

Release 1.17: May 2020

API:

Improved accessors for markup/ink/polygon annotation data.
Chapter based API for faster EPUB loading.
Add more documentation to header files.
Improved digital signature signing and verification.
Validate changes in a signed PDF file.

Build:

Moved windows build to VS2019 solution.

PDF:

Redaction now works on images and links as well as text.
Greek, Cyrillic, Chinese, Japanese, and Korean scripts in forms and annotations.
File attachment annotations.
Use CCITT Fax compression for 1-bit images when creating PDF files.

Changes to MuPDF since last meeting

Release 1.17: May 2020

EPUB:

More forgiving XHTML parsing.
Accelerator files to cache chapter data for faster EPUB loading.
Optimized memory use.

mutool run:

Edit Markup, Ink, and Polygon annotation data.
Fill out form fields.

viewer:

Ask for confirmation before closing a PDF with unsaved changes.
Embed and extract file attachment annotations.

Changes to MuPDF since last meeting

Release 1.18: September 2020

API:

- C++ and Python bindings.   Python bindings available with pip install mupdf
- De-hyphenation option in structured text extraction.
- Added methods for pdf annotation/signature date management.
- Added choice of image redaction algorithms: none, full, partial.
- Optional use of Tesseract to use OCR to extract text.

mupdf-gl:

- Added IBM Common User Access shortcuts for copy & paste.
- Improved redaction UI.

Changes to MuPDF since last meeting

Release 1.18: September 2020

PDF:

High security redaction -- save redacted PDF as flattened bitmap, guaranteed to not leak any sensitive redacted information.

Java :

Example desktop Java viewer.

HTML5:

Added parser.

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)

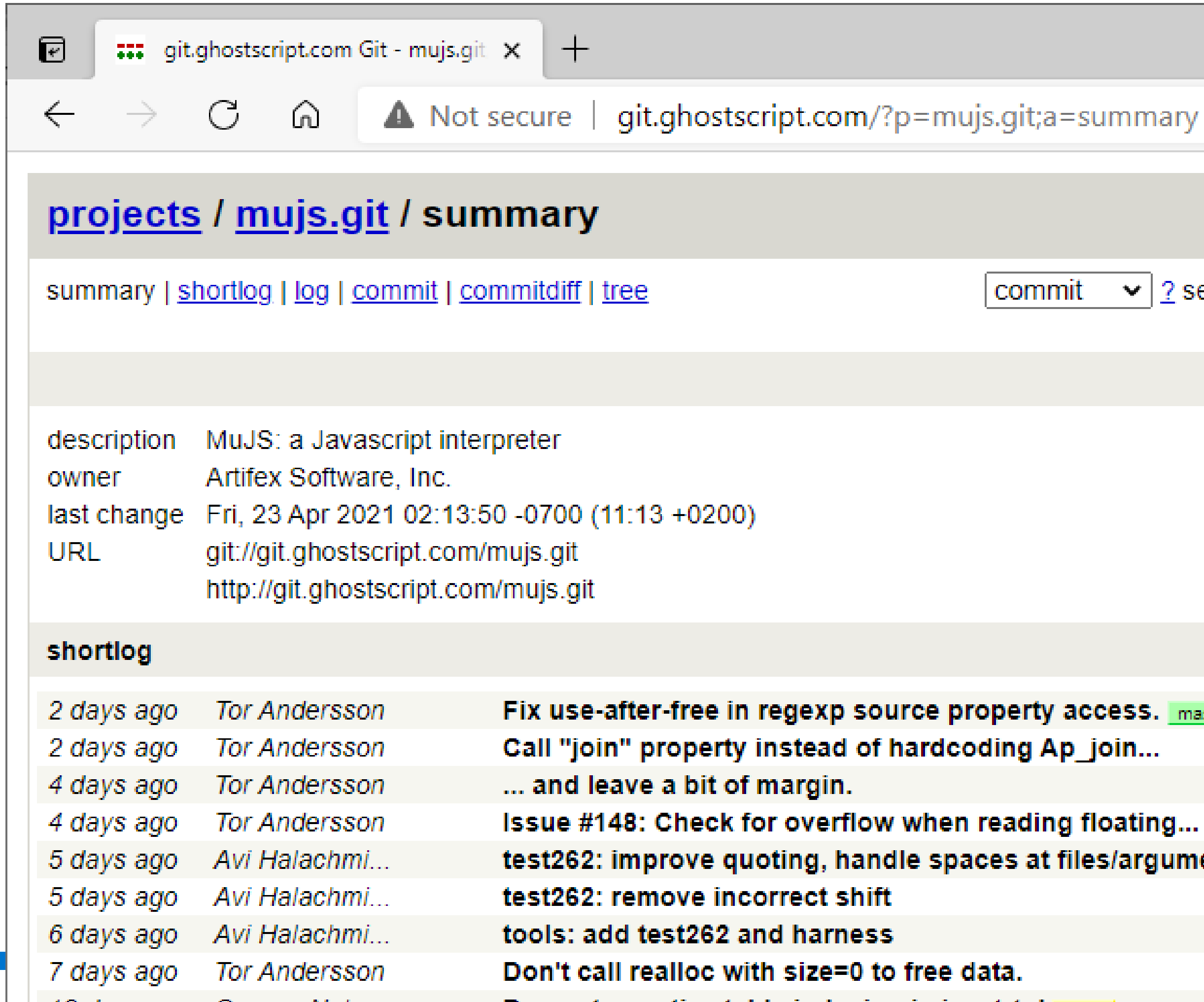
MuJS

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


git.ghostscript.com Git - mujs.git
Not secure
git.ghostscript.com/?p=mujs.git;a=summary

projects / mujs.git / summary

summary | shortlog | log | commit | commitdiff | tree

description	MuJS: a Javascript interpreter
owner	Artifex Software, Inc.
last change	Fri, 23 Apr 2021 02:13:50 -0700 (11:13 +0200)
URL	git://git.ghostscript.com/mujs.git http://git.ghostscript.com/mujs.git

shortlog

2 days ago	*Tor Andersson*	**Fix use-after-free in regexp source property access.**
2 days ago	*Tor Andersson*	**Call "join" property instead of hardcoding Ap_join...**
4 days ago	*Tor Andersson*	**... and leave a bit of margin.**
4 days ago	*Tor Andersson*	**Issue #148: Check for overflow when reading floating...**
5 days ago	*Avi Halachmi...*	**test262: improve quoting, handle spaces at files/argume**
5 days ago	*Avi Halachmi...*	**test262: remove incorrect shift**
6 days ago	*Avi Halachmi...*	**tools: add test262 and harness**
7 days ago	*Tor Andersson*	**Don't call realloc with size=0 to free data.**

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.

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)

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

New PDF interpreter in Ghostscript. Significant progress. See pdfi branch in repository

Ghostscript device interface cleanup:
https://www.notion.so/The-Great-Device-Rework-Of-2021-94092fe1395d4a088b91462f0ca5038a

Resolution independent display list for Ghostscript.

Digital signatures.

Docx output for Ghostscript.

More Information

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

Ghostscript discussions on IRC freenode #ghostscript channel
MuPDF discussions on IRC freenode #mupdf channel

Bug reports
bugs.ghostscript.com

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