Collabora Productivity

PDFium for better PDF rendering and editing

By Ashod Nakashian

Consultant Software Engineer at Collabora Productivity

ashod.nakashian@collabora.com

Overview

- PDF: de facto standard of print-accurate documents

- The problem in picturesss

- PDFium: accurate, fast, compatible

- Challenges

- Results

The Problem In Pictures

The Problem In Pictures

Greetings and Set Phrases

おはようございます。
Ohayoo gozaimasu.
Good morning.

こんにちは。
Konnichiwa.
Hello.

こんばんは。
Konbanwa.
Good evening.

お元気ですか。 Ogenki desu ka.
Are you well?

はい、おかげさまで。 Hai, okage-sama de.
Yes, thank you.

お疲れさまでした。
Otsukare-sama deshita.
You have worked hard.
(Set phrase when finishing)work

お先に失礼します。
Osaki ni shitsuree-shimasu.
Excuse me for leaving first.
(Set phrase when leaving)work

ありがとう［ございます］。
Arigatou gozaimasu.
Thank you very much.

どういたしまして。
Doo itashimashite.
You're welcome.
いいえ。
Iie.
Not at all.

［どうも］すみません。
[Doomo] sumimasen.
Excuse me.
ごめんなさい。
Gomennasai.
I'm sorry.

いいえ。 Iie.
Not at all.

The Problem In Pictures

1948.12.10

 !"#$%&'()#&*+,-.#/01(
)#2345678(9:3;<=>$?@A&B(&3

-CD=>EFG3HI/JK6LMNOPQR/SNTU3VW=>XY
9:GZ8T[3\]=>^_+,45`aG3Ycdefgh#U
iVj[N&3

kGlm#no#6p(eqrs#Utu6vw()#G+,xy6(Nz
623{|}6xM-/
~ﾑ()#Gﾒ·&B(&3
- k□□□□□/□□()#Gﾒ·&B(&3

□□- □23 ﺩﬡﴬﭏﴢﭏﬓﭏﬕ-
ﬦשׂﬕﬔﴬﬧﭏ□□□□ﬦﬓ□□□3Y□□*+9:y□&□□ #¡¢£
¤¥¦#/□□()#/§¨UN&3

©ª 23 □□#«¬U3-=>@□□9:®□+!¯=>°±□□/
□/□#/ﾳ´UN&3

Ashod Nakashian @ LibreCon2018, Tirana

26/09/18

www.collaboraoffice.com

Challenges

- PDF is print-format, not document-format

- Designed to reproduce pixels in correct position on page

- Editing context lost when PDF produced

  - In many cases, each character is positioned individually

  - No spaces between works

  - No new-lines

Collabora Productivity

www.collaboraoffice.com

As Displayed

Greetings and Set Phrases

Editable

Greetings and Set Phrases

Why Poppler *not*?

- Poor / missing support for Asian scripts

- Incompatible license => Costly out-of-process wrapper (see next point)

- Poor performance due to wrapper

  - PDF => text file => parser

- Sub-par accuracy

www.collaboraoffice.com

Immediate goals

- View-only PDF documents in Online

- Avoid creating slow and memory-hungry editable draw objects/shapes

- Avoid inaccuracies due to visual editing elements (such as space around text and difficulty of perfectly placing each character).

- Available on desktop by defining the environment variable LO_IMPORT_USE_PDFIUM

www.collaboraoffice.com

Why PDFium?

- Fast parsing

- Fast and accurate bitmap rendering

- Has heuristics to restore both spaces and new-lines

  - Helps support editing text and formating

- Mature API

- Very wide deployment

  - e.g. Chromium

- Compatible license

Collabora Productivity

www.collaboraoffice.com

Phase I: PDFium => Raster (Completed)

- Use PDFium to very quickly and accurately render PDF pages to bitmap images (stored as PNG in memory and on disk).

- Useful for view-only mode in Online

- Save both time by rasterizing directly

  - Can saves memory too, but sometimes it bloats

- Can use embedded glyphs for scripts instead of requiring fonts on system

- Use break command on rendered image to convert to editable objects / shapes

Collabora Productivity

www.collaboraoffice.com

Phase I: Implementation

- Existing PDF-to-Bitmap wrapper extended for multiple pages and page ranges

- Many API functions to export PDF structure details missing from PDFium:

  - 32 new public functions implemented in PDFium

  - Most all upstreamed

- New importing class: SdPdfFilter

- Original PDF streams stored with each rasterized page (to support editing) are now shared and stored once

- Support for swapping of PDF and rendered images

www.collaboraoffice.com

Phase II: Replace Poppler (TBD)

- Implement any missing API functionality in PDFium

- Improve the positioning accuracy of objects/shapes

- Implement complex shape properties, such as:

  - Bevelled edges

  - Line endings

  - Self-intersecting shapes

- Test on a much larger population of documents to catch major regressions

- Add / improve unit-tests

Collabora Productivity

www.collaboraoffice.com

Results: Before & After

5 DAY

JUMP-START

AND GUIDE TO MINI-FASTING™

See what a difference mini-fasting can make! With this powerful tool, you get all the health benefits of a fast … without feeling hungry!

What Is the 5-Day Mini-Fast™ JUMP-START?

I was a bodybuilder and a fitness fanatic in my twenties, and I even earned the title of "Ms. St. Louis" after winning one bodybuilding competition. I followed the Arnold Classic every year, and adopted a bodybuilding diet regimen from which my trainers and I never deviated.

On this regimen, every two hours was feeding time, and it was the golden rule of bodybuilding that you never, ever miss a meal. I had the alarm on my sports watch set to go off every two hours so I could stuff that boiled chicken down my throat.

Well, I've come a long way, baby! These days, I know that fasting is a far better fitness and weight-loss strategy than grazing. **I've studied the science, I've lived the results, and I've helped thousands of other people get fit and slim by fasting.**

I've also discovered that going without food once in awhile is not going to kill me. **In fact, it makes me feel — and look — better!**

Now, there are all kinds of fasts, and I'm not talking about fasts in which you drink nothing but water or juice. I don't recommend that type of fasting. In the type of fast I'm talking about, you simply restrict the amount of time you're allowed to eat during the day and focus on eating simple, healthy foods.

HOW DOES A MINI-FAST HELP MAXIMIZE WEIGHT LOSS?

- It minimizes your calories, so it gives your body no choice but to burn stored fat for energy.

- It helps you balance your blood sugar, because it improves your insulin response to blood sugar spikes.

My House

My house is a colonial-style villa in a noble residential area of San Francisco, with a magnificent view of the Bay. Originally, the facade of the house was painted in a light shade of yellow, but today, the side facing the street is now overgrown with wild roses up to the second floor windows. As you approach the house from the _____ comes into view. In the carp_____ timer, still in very good conditi_____ urely drives along the coast.

First Floor

The front door opens to a tiled _____ than a coat rack and a few p_____ there isn't much else to see her_____

In the hallway, it begins to get _____ ing. Both walls are lined with _____ ern art from artists such as A_____ like. An extravagant lighting _____ the appropriate atmosphere. H_____ room, a 200 year old crystal m_____ stark contrast to the Pop-art.

Four tall archways open up to _____

Artists in the hallway:	
	Warhol
	Liechtenstein
	Hockney
	Mondrian

To demonstrate that there exists an infinite length game for all n, we must demonstrate the existence of a real, increasing, nonzero eigenvector/value pair $\mathbf{v_n}, \lambda_n$ for all n.

3.3. Establishing and bounding a positive real eigenvalue.

$$S_n = T_n - \lambda I_n = \begin{bmatrix} -1-\lambda & 1 & 0 & 0 & \ldots & 0 \\ 0 & -1-\lambda & 1 & 0 & \ldots & 0 \\ 0 & 0 & -1-\lambda & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & 0 & -1-\lambda & 1 \\ -1 & 0 & \ldots & \ldots & 0 & 1-\lambda \end{bmatrix}$$

Expanding $\det(S_n)$ by cofactors along the bottom row, we see

$$\det(S_n) = -1(-1)^{1+n}\begin{vmatrix} 1 & 0 & 0 & \ldots & 0 \\ -1-\lambda & 1 & 0 & \ldots & 0 \\ 0 & -1-\lambda & 1 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & -1-\lambda & 1 \end{vmatrix} +$$

$$(1-\lambda)(-1)^{n+n}\begin{vmatrix} -1-\lambda & 1 & 0 & \ldots & 0 \\ 0 & -1-\lambda & 1 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & \ldots & 0 & -1-\lambda \end{vmatrix}$$

The determinant in the first term reduces to 1, and the determinant in the second term reduces to $-(\lambda)^{n-1}$. The characteristic polynomial of T_n is then $(-1)^n + (1-\lambda)(-1)(-1-\lambda)^{n-1} = 0$. Expanding, we have

$$(-1)^{2+n} + (-1-\lambda)^{n-1} - \lambda(-1-\lambda)^{n-1} = 0$$

Diagnostic Usability Testing—St. Louis, Missouri

Notice Rotation

We used five scenarios for the testing in St. Louis. The scenarios determined which notice and style participants received first—a page 1; which two additional notices in that same style they received during Task 2—the introduction of page 2; and which alternate style for Task 3—table or prose. The first three scenarios (MO 501–MO 504) introduced a prose style first in Task 1 and 2, and participants compared the prose to the table in Task 3. In scenario 4 participants (MO 505 and MO 506) were presented with only the table versions alternating between the two styles for each participant. They did not see a prose version. For the last two participants (MO 507 and MO 508), the styles were rotated for the purpose of seeing how well participants were able to compare the sharing practices of the three banks across varying styles of presentation (table and prose) rather than across the same style.

For Task 1, participants worked with only page 1, in either the prose or the table format. For Task 2 of scenarios 1–4, participants compared the three banks' notices in that same table or prose format. For Task 2 of scenario 5, participants compared two table versions against a prose version. During the last section (Task 3), each participant was presented with the alternate table or prose format of the bank notice they worked with in Task 1.
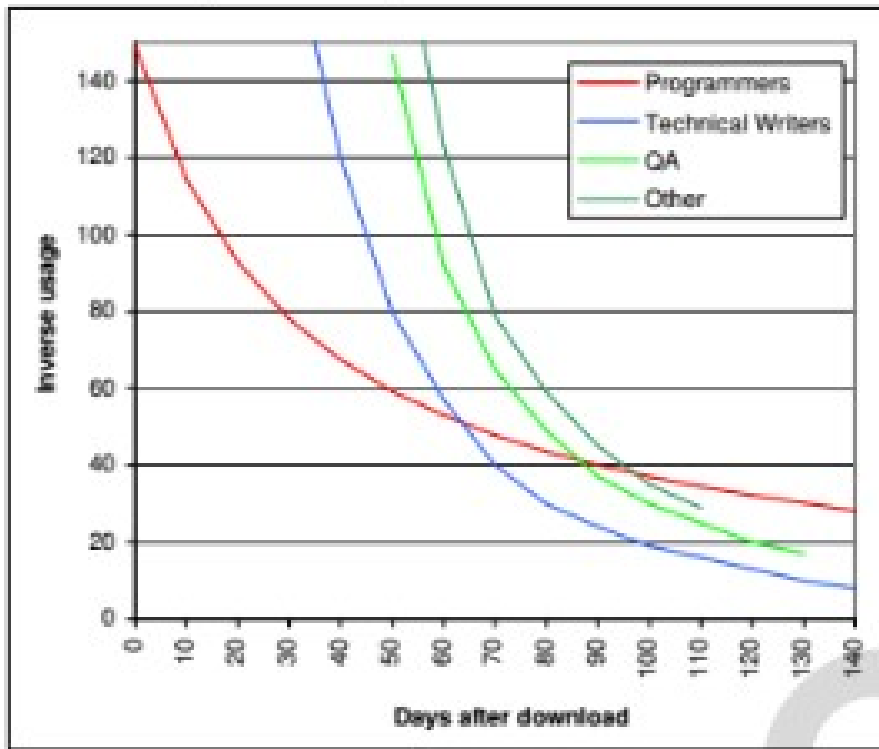
Rotation

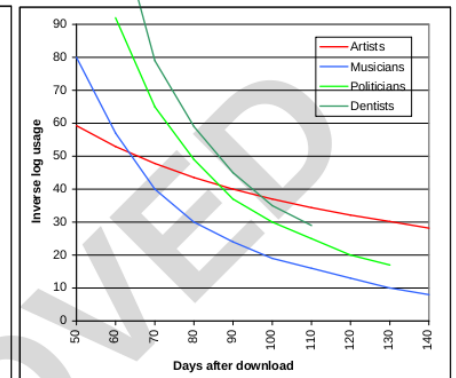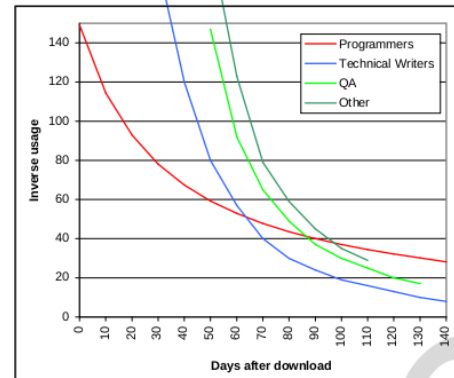| Participant Number	Scenario	Task 1: Think Aloud (1st notice participants received)	Task 3 Comparisons (2nd & 3rd notice participants received randomly*)		
MO 501	Scenario 1	Neptune prose (page 1)	Mercury & Mars prose (page 1)	Page 2 for three notices	Neptune, Mercury & Mars table version 1
MO 502	Scenario 2	Mercury prose (page 1)	Neptune & Mars prose (page 1)	Page 2 for three notices	Neptune, Mercury & Mars table version 1
MO 503	Scenario 3	Mars prose (page 1)	Neptune & Mercury (page 1)	Page 2 for three notices	Neptune, Mercury & Mars table version 1
MO 504	Scenario 1	Neptune prose (page 1)	Mercury & Mars prose (page 1)	Page 2 for three notices	Neptune, Mercury & Mars table version 1
MO 505	Scenario 4	Mars table version 1 (page 1)	Neptune & Mercury table	Page 2 for three notices	Neptune, Mercury & Mars
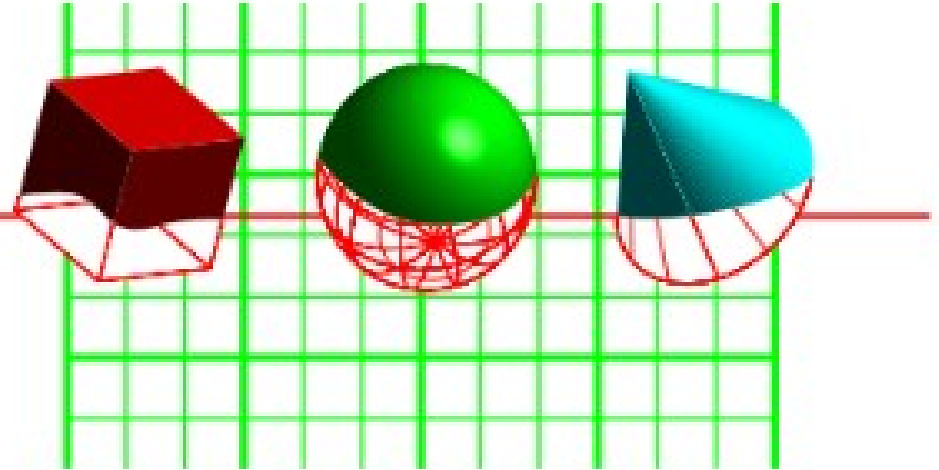
A Graphical Representation of Inverse VRML Uptake

Change the number in red below to adjust for download rate and/or bandwidth.
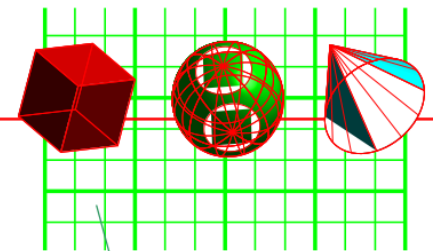
| 1 | The number 1 represents an engineer with an "average" cube *

A Graphical Representation of Inverse VRML Uptake

Change the number in red below to adjust for download rate and/or bandwidth.

| 1 | The number 1 represents an engineer with an "average" cube *
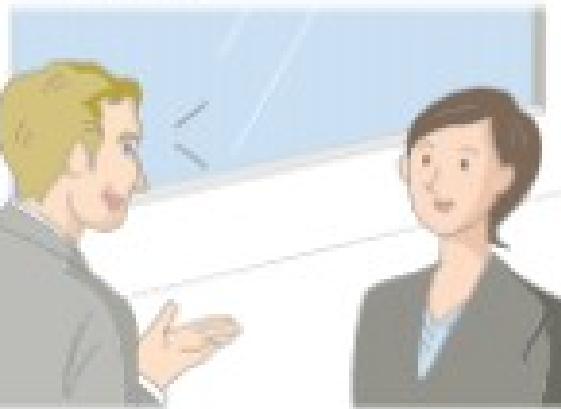
Greetings and Set Phrases

おはようございます。
Ohayoo gozaimasu.
Good morning.

こんにちは。
Konnichiwa.
Hello.

お元気ですか。 Ogenki desu ka.
Are you well?

はい、おかげさま
Yes, thank you.

Memory footprint

Average: 97.17%

Collabora Productivity

www.collaboraoffice.com

Collabora Productivity

<Your Question Here/>

By Ashod Nakashian

Ashod.nakashian@collabora.com

Collabora Productivity

Thank you!

By Ashod Nakashian

Ashod.nakashian@collabora.com