PDFBOX-5033

nuwanda

2020-12-04

Contents

Chapter 1

Root issue PDFBOX-5033

1.1 Summary

CFF FontParser exits with illegal offset in font

1.2 Description

Dear Devs,

we've encountered an issue with version 2.0.20 and 2.0.21 of PDFbox when trying to parse a PDF for text extraction that seem to have existed before see FOP-2751.

I reproduced this issue with the pdfbox-app and the FuturaStd-Book.pdf of FOP-2751:
{noformat:title=Console output}

```
java -jar pdfbox-app-2.0.21.jar ExtractText FuturaStd-Book.pdf
Dez 04, 2020 11:06:00 AM org.apache.pdfbox.pdmodel.font.PDType1CFont <init>
SCHWERWIEGEND: Can't read the embedded Type1C font FuturaStd-Book
java.io.IOException: illegal offset value 2949166 in CFF font
at org.apache.fontbox.cff.CFFParser.readIndexDataOffsets(CFFParser.java:192)
at org.apache.fontbox.cff.CFFParser.readIndexData(CFFParser.java:201)
at org.apache.fontbox.cff.CFFParser.parseFont(CFFParser.java:484)
at org.apache.fontbox.cff.CFFParser.parse(CFFParser.java:122)
at org.apache.fontbox.cff.CFFParser.parse(CFFParser.java:75)
at org.apache.pdfbox.pdmodel.font.PDType1CFont.<init>(PDType1CFont.java:102)
at org.apache.pdfbox.pdmodel.font.PDFontFactory.createFont(PDFontFactory.java:74)
at org.apache.pdfbox.pdmodel.PDResources.getFont(PDResources.java:146)
at org.apache.pdfbox.contentstream.operator.text.SetFontAndSize.process(SetFontAndSize.java:66)
at org.apache.pdfbox.contentstream.PDFStreamEngine.processOperator(PDFStreamEngine.java:933)
at org.apache.pdfbox.contentstream.PDFStreamEngine.processStreamOperators(PDFStreamEngine.java:515)
at org.apache.pdfbox.contentstream.PDFStreamEngine.processStream(PDFStreamEngine.java:489)
at org.apache.pdfbox.contentstream.PDFStreamEngine.processPage(PDFStreamEngine.java:156)
at org.apache.pdfbox.text.LegacyPDFStreamEngine.processPage(LegacyPDFStreamEngine.java:144)
at org.apache.pdfbox.text.PDFTextStripper.processPage(PDFTextStripper.java:397)
at org.apache.pdfbox.text.PDFTextStripper.processPages(PDFTextStripper.java:325)
at org.apache.pdfbox.text.PDFTextStripper.writeText(PDFTextStripper.java:272)
at org.apache.pdfbox.tools.ExtractText.extractPages(ExtractText.java:377)
at org.apache.pdfbox.tools.ExtractText.startExtraction(ExtractText.java:274)
at org.apache.pdfbox.tools.ExtractText.main(ExtractText.java:97)
at org.apache.pdfbox.tools.PDFBox.main(PDFBox.java:60)
```

```
Dez 04, 2020 11:06:00 AM org.apache.pdfbox.pdmodel.font.FileSystemFontProvider loadDiskCache
WARNUNG: New fonts found, font cache will be re-built
Dez 04, 2020 11:06:00 AM org.apache.pdfbox.pdmodel.font.FileSystemFontProvider <init>
WARNUNG: Building on-disk font cache, this may take a while
Dez 04, 2020 11:06:02 AM org.apache.pdfbox.pdmodel.font.FileSystemFontProvider <init>
WARNUNG: Finished building on-disk font cache, found 550 fonts
Dez 04, 2020 11:06:02 AM org.apache.pdfbox.pdmodel.font.PDType1CFont <init>
WARNUNG: Using fallback font Courier for FuturaStd-Book
```

{noformat}

Other examples fonts causing this issue are:
* Can't read the embedded Type1C font COGXUZ+MetaPlusNormal-Caps
* Can't read the embedded Type1C font DJTRFS+MetaPlusBold-CapsItalic
* Can't read the embedded Type1C font EAFTRP+MetaPlusNormal-Caps
* Can't read the embedded Type1C font GQHJVM+MetaPlusNormal-CapsItalic
* Can't read the embedded Type1C font GUEVYR+MetaPlusBold-CapsItalic
* Can't read the embedded Type1C font HYTBMP+MetaPlusNormal-CapsItalic
* Can't read the embedded Type1C font IJCQXI+MetaPlusMedium-Caps
* Can't read the embedded Type1C font JRIYJF+MetaPlusNormal-Caps
* Can't read the embedded Type1C font JSQSJF+NeuzeitGro-Reg
* Can't read the embedded Type1C font KUZTXD+MetaPlusBook-Roman
* Can't read the embedded Type1C font LWIPLB+1496148105355.00001Arial.000-1
* Can't read the embedded Type1C font MCDJBA+MetaSerif-BoldIta
* Can't read the embedded Type1C font UNLUJK+Barmeno-Medium

I couldn't find another issue about this. Is this already known?

1.3 Attachments

No attachments

1.4 Comments

1. **msahyoun:** The fop sample document you're referring to has been generated before or after the fix was done in fop? FuturaStd-Book_full.pdf works fine.

2. **tilman:** The file mentioned ([FuturaStd-Book.pdf|https://issues.apache.org/jira/secure/attachment/12891693/Fut
Book.pdf]) is broken. The exception message isn't really good, because the actual problem happens a bit earlier. I'll fix that but it won't change anything.

3. **jira-bot:** Commit 1884105 from Tilman Hausherr in branch 'pdfbox/branches/2.0'
[https://svn.apache.org/r1884105]

   PDFBOX-5033: throw exception on illegal offSize value

4. **jira-bot:** Commit 1884106 from Tilman Hausherr in branch 'pdfbox/branches/1.8'
[https://svn.apache.org/r1884106]

   PDFBOX-5033: throw exception on illegal offSize value

5. **jira-bot:** Commit 1884107 from Tilman Hausherr in branch 'pdfbox/trunk'
[https://svn.apache.org/r1884107]

PDFBOX-5033: throw exception on illegal offSize value

6. **tilman:** There's a better exception text now. My understanding is that the file is the result of a FOP bug. If you see this differently, please comment and/or reopen.

Chapter 2

Connected issue FOP-2751

2.1 Summary

Acrobat Reader error with some Latin Fonts

2.2 Description

Some Latin Fonts cannot be embedded into PDF document.
How to repeat
1. Get FOP from trunk@1811797
2. Get Font from the attachment
3. Use my config file and generate the PDF, there is no error reported.
4. Open the PDF file in Acro-Reader, it will report the "cannot extract the embedded font" error.

CFFParser@FuturaStd-Book.fo

```
java.io.IOException: illegal offset value 2949166 in CFF font
        at org.apache.fontbox.cff.CFFParser.readIndexDataOffsets(CFFParser.java:188)
        at org.apache.fontbox.cff.CFFParser.readIndexData(CFFParser.java:197)
        at org.apache.fontbox.cff.CFFParser.parseFont(CFFParser.java:460)
        at org.apache.fontbox.cff.CFFParser.parse(CFFParser.java:145)
```

GillSansStd.fo and TimesNewRomanMTStd-Bold.fo detects no Error in CFFParser.
I attached the result of checking the CFF structure in Acrobat.
[^GillSansStd_Acrobat.png]
[^TimesNewRomanMTStd-Bold_Acrobat.png]

2.3 Attachments

1. fop.xconf
2. FuturaStd-Book_full.pdf
3. FuturaStd-Book.fo
4. FuturaStd-Book.pdf
5. GillSansStd_Acrobat.png

6. GillSansStd__full.pdf
7. GillSansStd.fo
8. GillSansStd.pdf
9. TimesNewRomanMTStd-Bold__Acrobat.png
10. TimesNewRomanMTStd-Bold__full.pdf
11. TimesNewRomanMTStd-Bold.fo
12. TimesNewRomanMTStd-Bold.pdf

2.4 Comments

1. **ssteiner:** http://svn.apache.org/viewvc?view=revision&revision=1811970
2. **murakami@brainsellers.com:** Thanks for fixing.