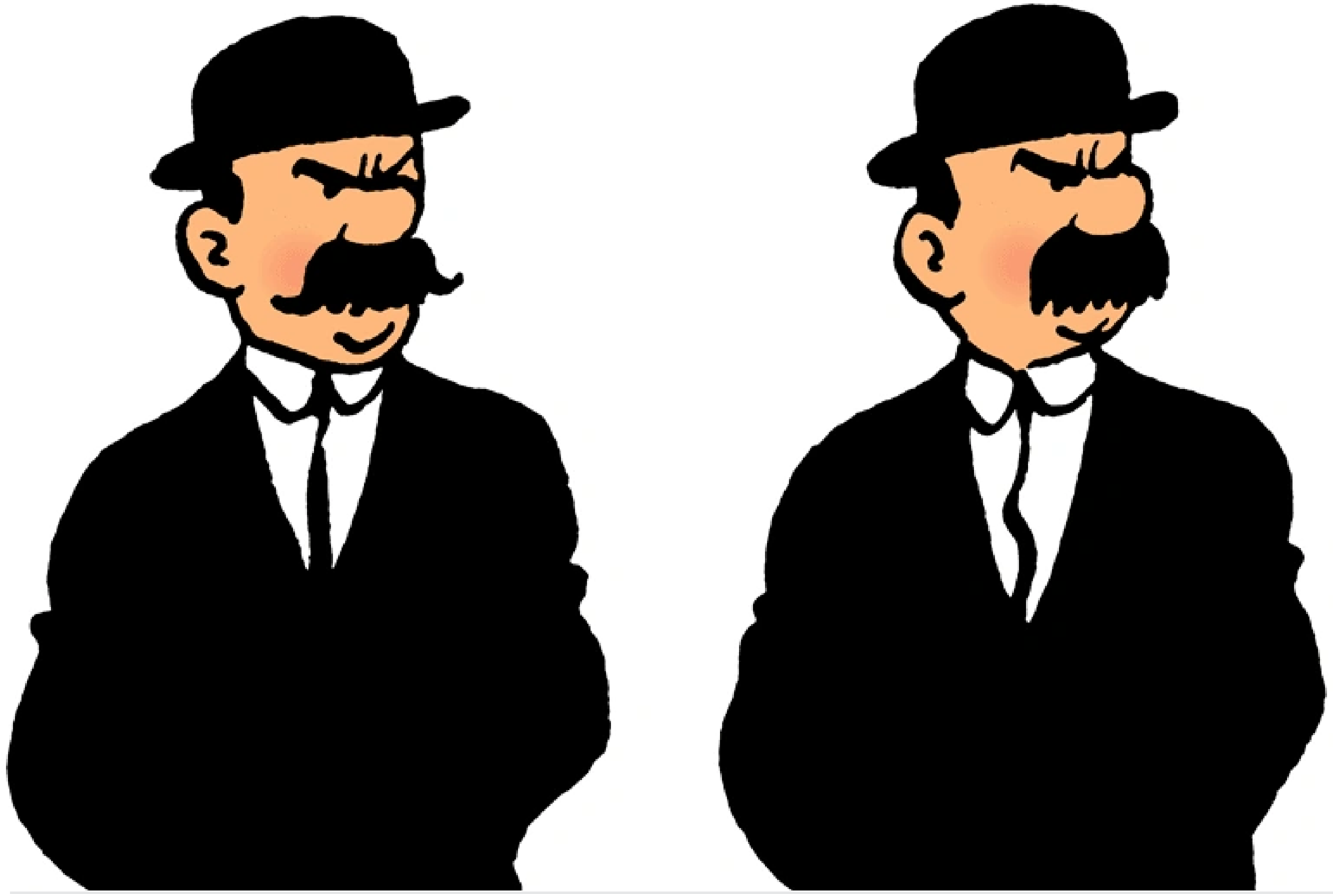


working together!

PDF Days Europe 2022 | Berlin

The PDF Detectives

The dynamically typed duo is on the case!

Who are we?

- Kevin Willems
  - Presales Engineer @ iText

- Michaël Demey
  - Research Manager @iText

Introducing iText Software

SDK

iText 7 Suite

Best documented and most feature-rich PDF library

Open-source library for PDF generation and management,

- PDF/A, PDF/UA,
- Digital signing,
- Security,
- HTML conversion,
- Redaction, OCR, etc.,

and closed source add-ons for rendering capabilities, advanced data extraction, advanced forms processing, PDF optimization, advanced language script capabilities, Office to PDF conversion and much more.

High convenience tools

iText DITO

Fast template-based document generation

High convenience technology block – save time
Based on iText technology

- **WYSIWYG Template designer**
- **Conditional logic for content visibility**
- **Data binding**
- **Advanced data visualizations (barcodes/QR codes, charts…)**
- **PDF/UA Compliance Assistance**
- **Template Manager with version, dependency and permission control**
- **Containerized RESTful PDF-producer API (also available as native Java SDK).**

Mission Briefing

Goal of the Presentation

- Talk about some interesting behavior in PDF files
  - Debug
  - Explain
  - Resolve (hopefully ☺)

The Toolbelt

Inventory Report!

working together!

Analysis and Debugging Tools

- iText RUPS (+iText Core)
- Acrobat Pro

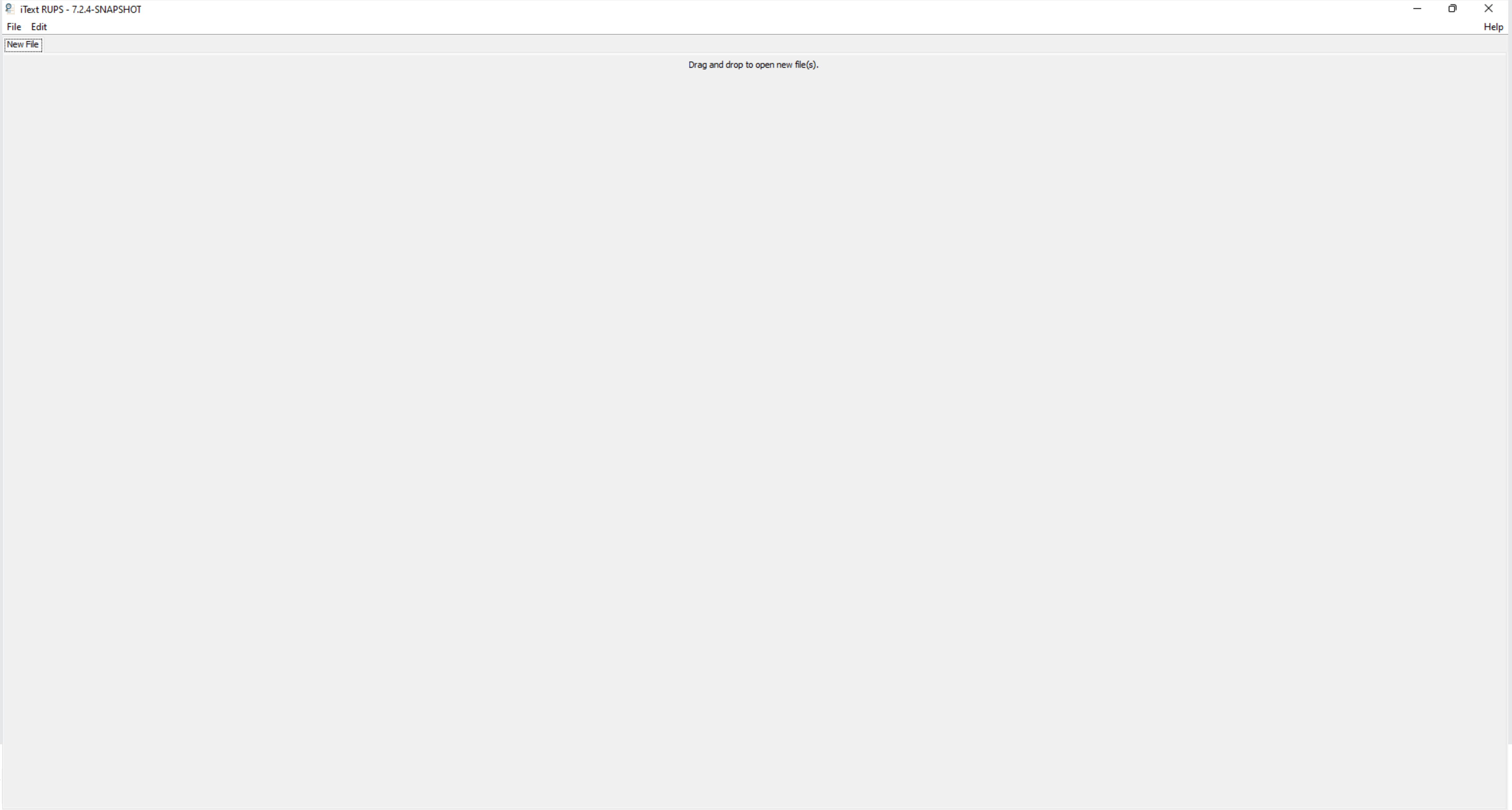
Accessibility and Tagging

- VeraPDF
- PAC

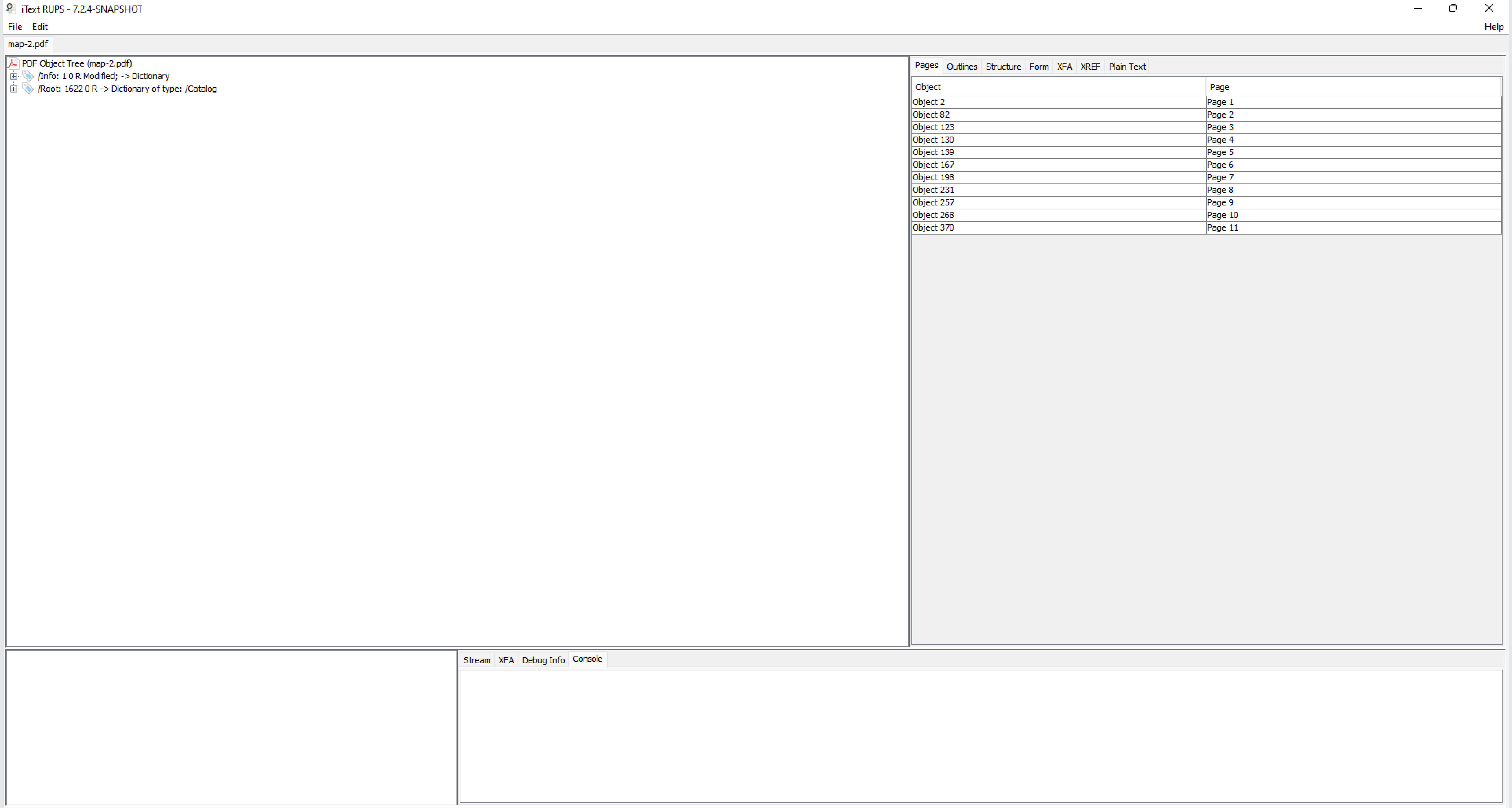
Low Level Tools

- Hex Editor
- Notepad++

RUPS “Demo”

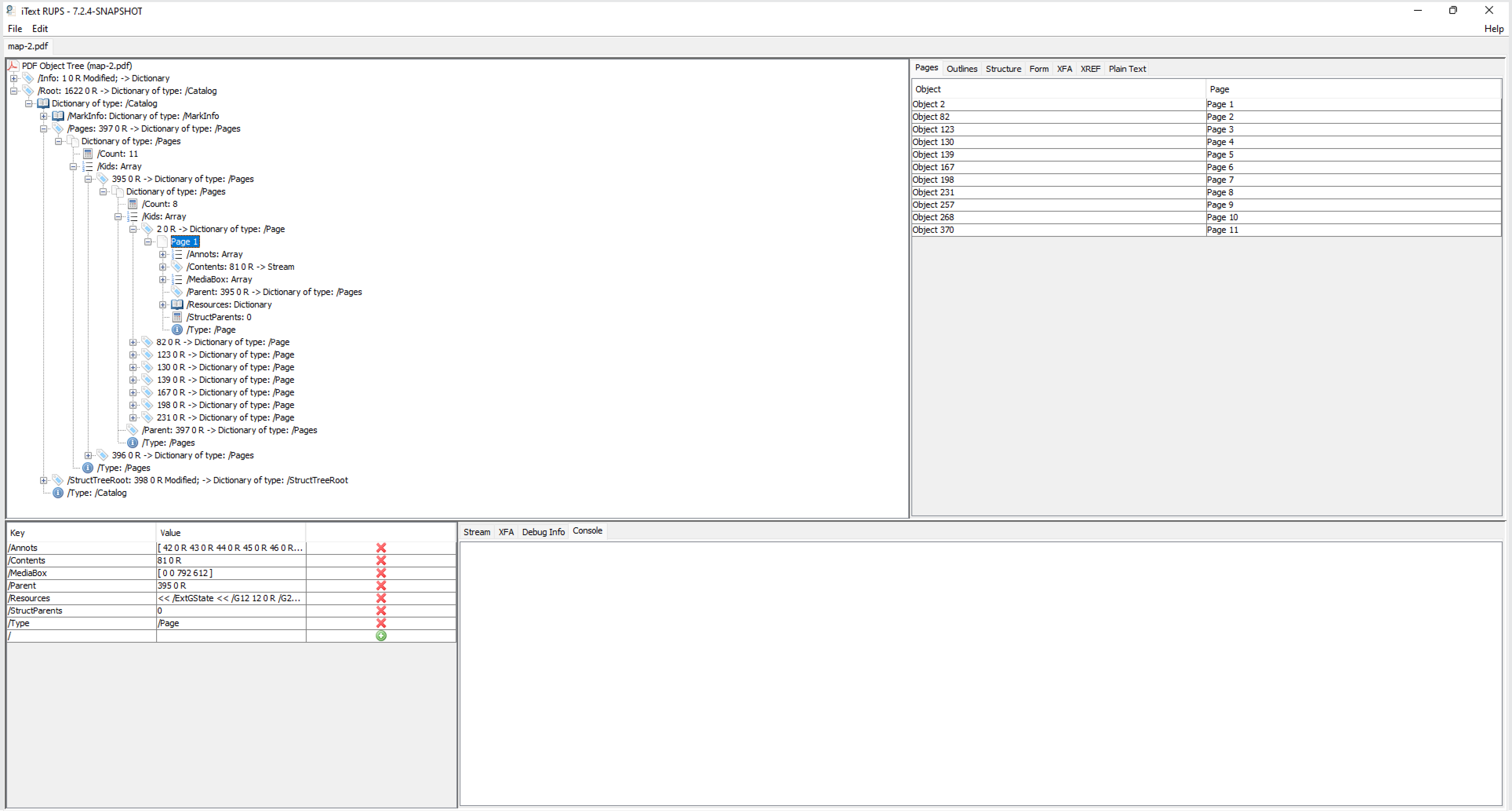


Drag and drop to open new file(s).



```
PDF Object Tree (map-2.pdf)
  /Info: 1 0 R Modified; -> Dictionary
  /Root: 1622 0 R -> Dictionary of type: /Catalog
```

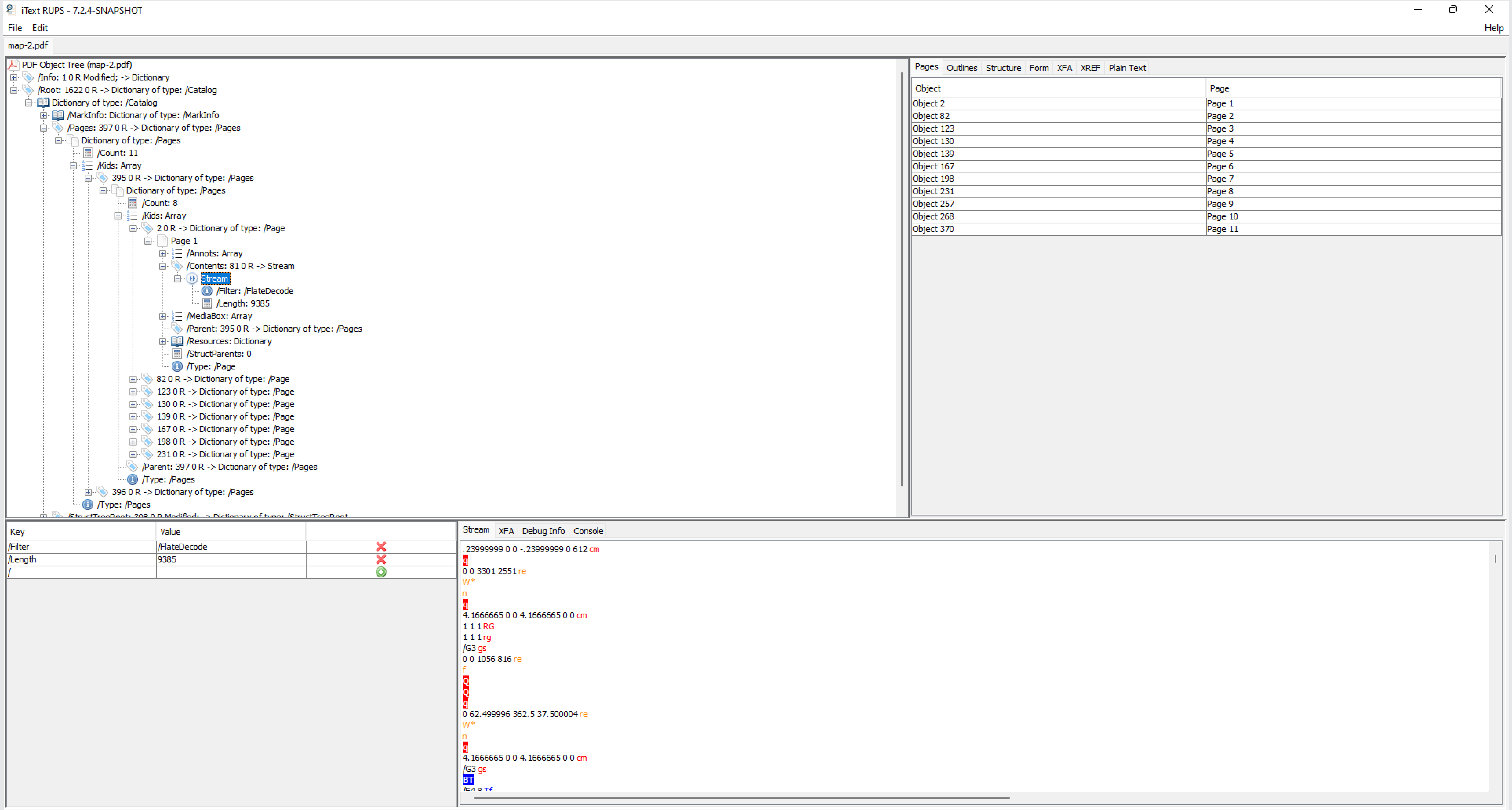
Object	Page
Object 2	Page 1
Object 82	Page 2
Object 123	Page 3
Object 130	Page 4
Object 139	Page 5
Object 167	Page 6
Object 198	Page 7
Object 231	Page 8
Object 257	Page 9
Object 268	Page 10
Object 370	Page 11



```
PDF Object Tree (map-2.pdf)
  /Info: 1 0 R Modified; -> Dictionary
  /Root: 1622 0 R -> Dictionary of type: /Catalog
    Dictionary of type: /Catalog
      /MarkInfo: Dictionary of type: /MarkInfo
      /Pages: 397 0 R -> Dictionary of type: /Pages
        Dictionary of type: /Pages
          /Count: 11
          /Kids: Array
            395 0 R -> Dictionary of type: /Pages
              Dictionary of type: /Pages
                /Count: 8
                /Kids: Array
                  2 0 R -> Dictionary of type: /Page
                    Page 1
                      /Annots: Array
                      /Contents: 81 0 R -> Stream
                      /MediaBox: Array
                      /Parent: 395 0 R -> Dictionary of type: /Pages
                      /Resources: Dictionary
                      /StructParents: 0
                      /Type: /Page
                  82 0 R -> Dictionary of type: /Page
                  123 0 R -> Dictionary of type: /Page
                  130 0 R -> Dictionary of type: /Page
                  139 0 R -> Dictionary of type: /Page
                  167 0 R -> Dictionary of type: /Page
                  198 0 R -> Dictionary of type: /Page
                  231 0 R -> Dictionary of type: /Page
                /Parent: 397 0 R -> Dictionary of type: /Pages
                /Type: /Pages
            396 0 R -> Dictionary of type: /Pages
          /Type: /Pages
      /StructTreeRoot: 398 0 R Modified; -> Dictionary of type: /StructTreeRoot
      /Type: /Catalog
```

Object	Page
Object 2	Page 1
Object 82	Page 2
Object 123	Page 3
Object 130	Page 4
Object 139	Page 5
Object 167	Page 6
Object 198	Page 7
Object 231	Page 8
Object 257	Page 9
Object 268	Page 10
Object 370	Page 11

Key	Value	
/Annots	[42 0 R 43 0 R 44 0 R 45 0 R 46 0 R...	✗
/Contents	81 0 R	✗
/MediaBox	[0 0 792 612]	✗
/Parent	395 0 R	✗
/Resources	<< /ExtGState << /G12 12 0 R /G2...	✗
/StructParents	0	✗
/Type	/Page	✗
/		⊕

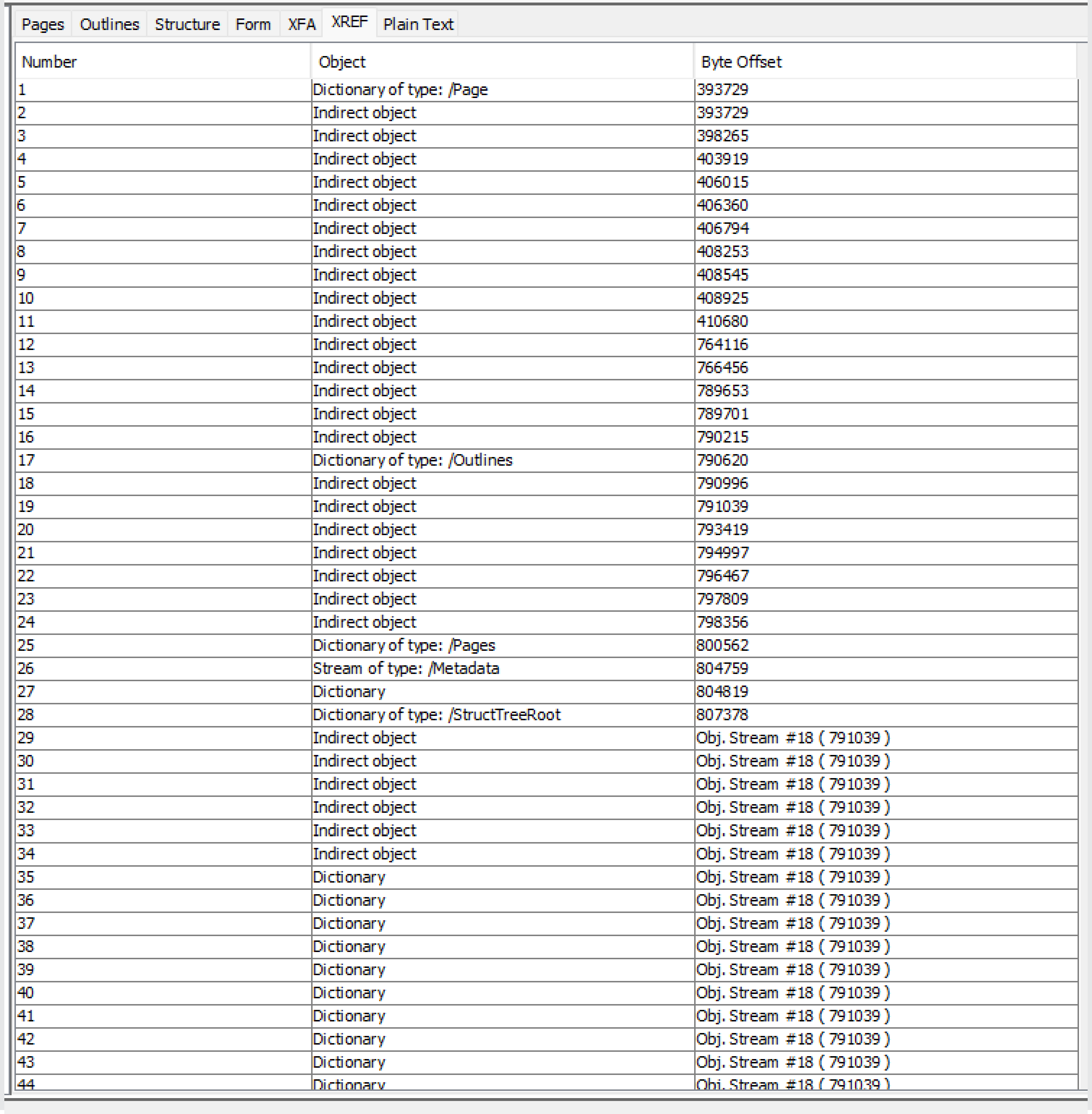


```
PDF Object Tree (map-2.pdf)
  /Info: 1 0 R Modified; -> Dictionary
  /Root: 1622 0 R -> Dictionary of type: /Catalog
    Dictionary of type: /Catalog
      /MarkInfo: Dictionary of type: /MarkInfo
      /Pages: 397 0 R -> Dictionary of type: /Pages
        Dictionary of type: /Pages
          /Count: 11
          /Kids: Array
            395 0 R -> Dictionary of type: /Pages
              Dictionary of type: /Pages
                /Count: 8
                /Kids: Array
                  2 0 R -> Dictionary of type: /Page
                    Page 1
                      /Annots: Array
                      /Contents: 81 0 R -> Stream
                        Stream
                          /Filter: /FlateDecode
                          /Length: 9385
                      /MediaBox: Array
                      /Parent: 395 0 R -> Dictionary of type: /Pages
                      /Resources: Dictionary
                      /StructParents: 0
                      /Type: /Page
                  82 0 R -> Dictionary of type: /Page
                  123 0 R -> Dictionary of type: /Page
                  130 0 R -> Dictionary of type: /Page
                  139 0 R -> Dictionary of type: /Page
                  167 0 R -> Dictionary of type: /Page
                  198 0 R -> Dictionary of type: /Page
                  231 0 R -> Dictionary of type: /Page
                /Parent: 397 0 R -> Dictionary of type: /Pages
                /Type: /Pages
            396 0 R -> Dictionary of type: /Pages
          /Type: /Pages
      /StructTreeRoot: 398 0 R Modified; -> Dictionary of type: /StructTreeRoot
```

Object	Page
Object 2	Page 1
Object 82	Page 2
Object 123	Page 3
Object 130	Page 4
Object 139	Page 5
Object 167	Page 6
Object 198	Page 7
Object 231	Page 8
Object 257	Page 9
Object 268	Page 10
Object 370	Page 11

Key	Value	
/Filter	/FlateDecode	✗
/Length	9385	✗
/		⊕

```
.23999999 0 0 -.23999999 0 612 cm
q
0 0 3301 2551 re
W*
n
q
4.1666665 0 0 4.1666665 0 0 cm
1 1 1 RG
1 1 1 rg
/G3 gs
0 0 1056 816 re
f
Q
Q
q
0 62.499996 362.5 37.500004 re
W*
n
q
4.1666665 0 0 4.1666665 0 0 cm
/G3 gs
BT
/F4 8 Tf
```



Number	Object	Byte Offset
1	Dictionary of type: /Page	393729
2	Indirect object	393729
3	Indirect object	398265
4	Indirect object	403919
5	Indirect object	406015
6	Indirect object	406360
7	Indirect object	406794
8	Indirect object	408253
9	Indirect object	408545
10	Indirect object	408925
11	Indirect object	410680
12	Indirect object	764116
13	Indirect object	766456
14	Indirect object	789653
15	Indirect object	789701
16	Indirect object	790215
17	Dictionary of type: /Outlines	790620
18	Indirect object	790996
19	Indirect object	791039
20	Indirect object	793419
21	Indirect object	794997
22	Indirect object	796467
23	Indirect object	797809
24	Indirect object	798356
25	Dictionary of type: /Pages	800562
26	Stream of type: /Metadata	804759
27	Dictionary	804819
28	Dictionary of type: /StructTreeRoot	807378
29	Indirect object	Obj. Stream #18 (791039)
30	Indirect object	Obj. Stream #18 (791039)
31	Indirect object	Obj. Stream #18 (791039)
32	Indirect object	Obj. Stream #18 (791039)
33	Indirect object	Obj. Stream #18 (791039)
34	Indirect object	Obj. Stream #18 (791039)
35	Dictionary	Obj. Stream #18 (791039)
36	Dictionary	Obj. Stream #18 (791039)
37	Dictionary	Obj. Stream #18 (791039)
38	Dictionary	Obj. Stream #18 (791039)
39	Dictionary	Obj. Stream #18 (791039)
40	Dictionary	Obj. Stream #18 (791039)
41	Dictionary	Obj. Stream #18 (791039)
42	Dictionary	Obj. Stream #18 (791039)
43	Dictionary	Obj. Stream #18 (791039)
44	Dictionary	Obj. Stream #18 (791039)

iText RUPS

- It’s on GitHub ☺

Inventory Report!

working together!

Analysis and Debugging Tools

- iText RUPS (+iText Core)
- Acrobat Pro

Accessibility and Tagging

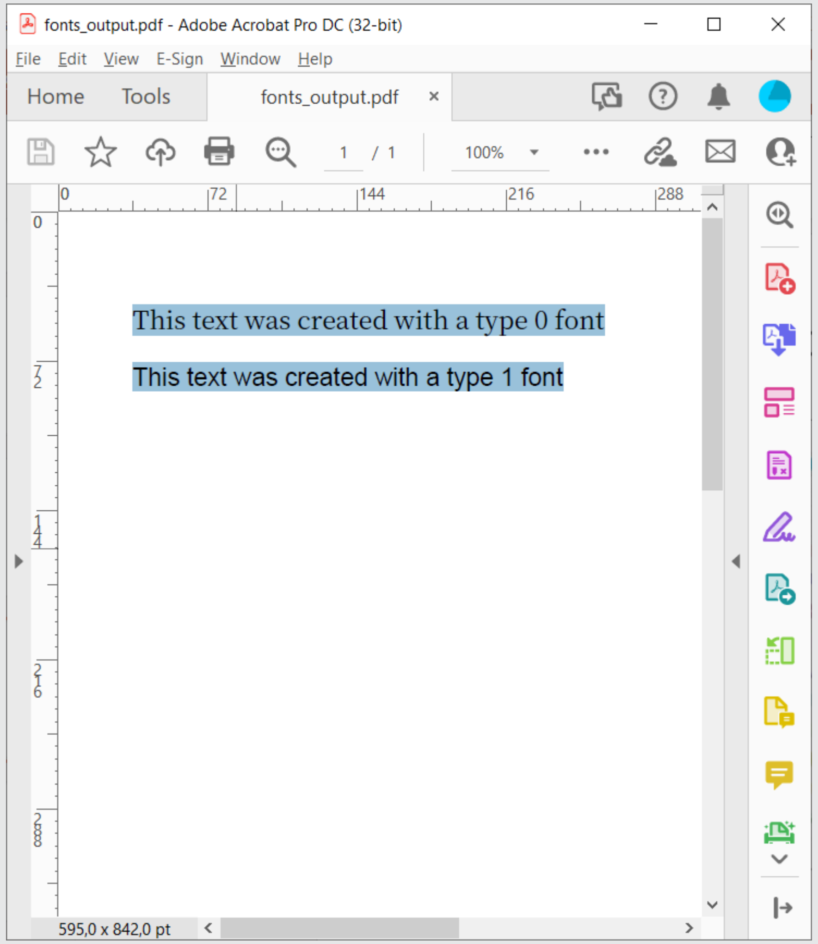
- VeraPDF
- PAC

Low Level Tools

- Hex Editor
- Notepad++

Cases

Case 1 - Usecase


fonts_output.pdf - Adobe Acrobat Pro DC (32-bit)
File Edit View E-Sign Window Help
Home Tools fonts_output.pdf ×
1 / 1
100%
0 72 144 216 288
595,0 x 842,0 pt

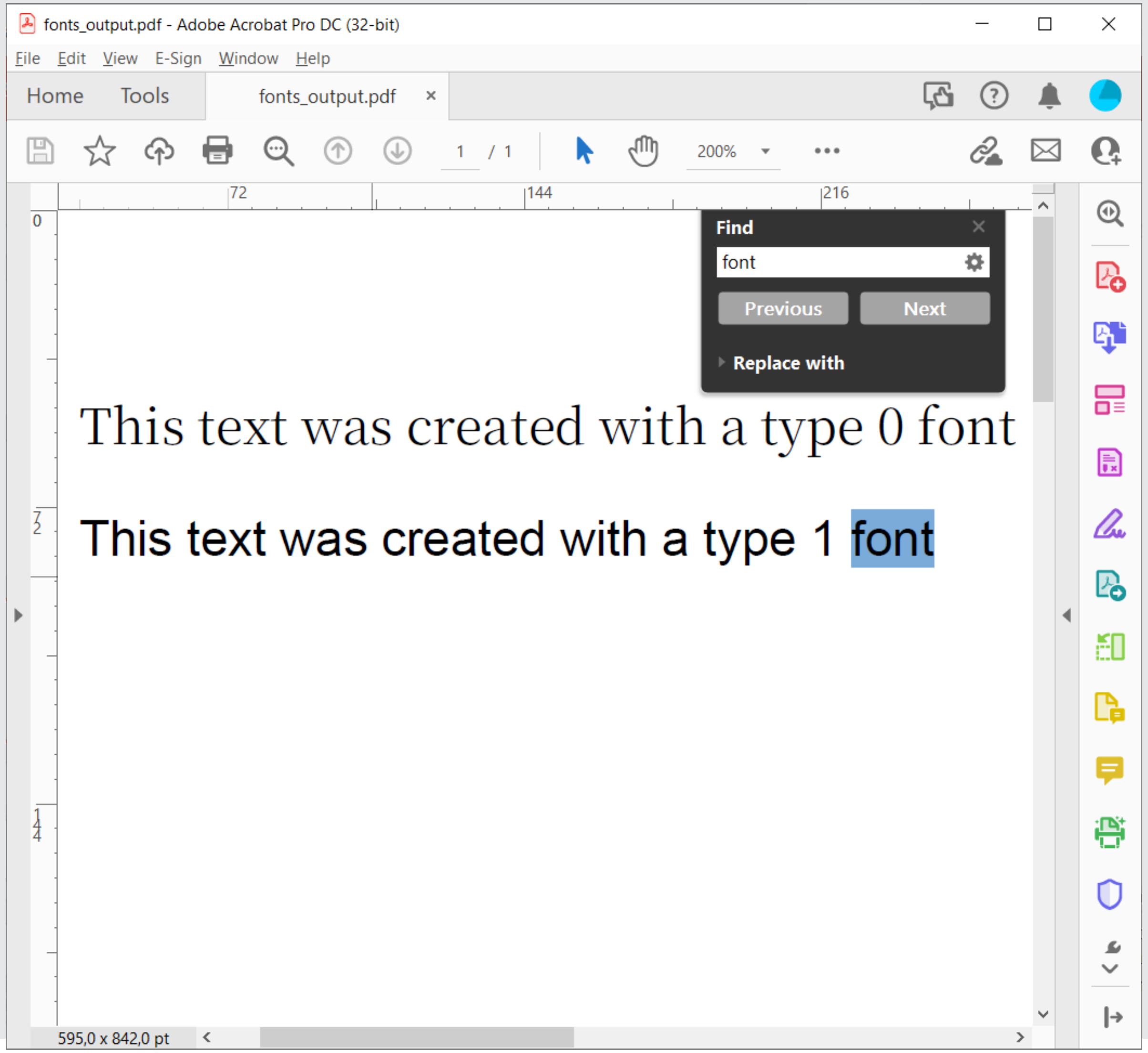
This text was created with a type 0 font

This text was created with a type 1 font

→

☒☒☒☒☒☒☒☒☒☒☒☒☒☒☒☒☒☒☒☒☒
☒☒☒☒☒☒☒☒☒☒☒☒☒☒☒☒☒☒☒
This text was created with a type 1 font

Case 1 - Usecase

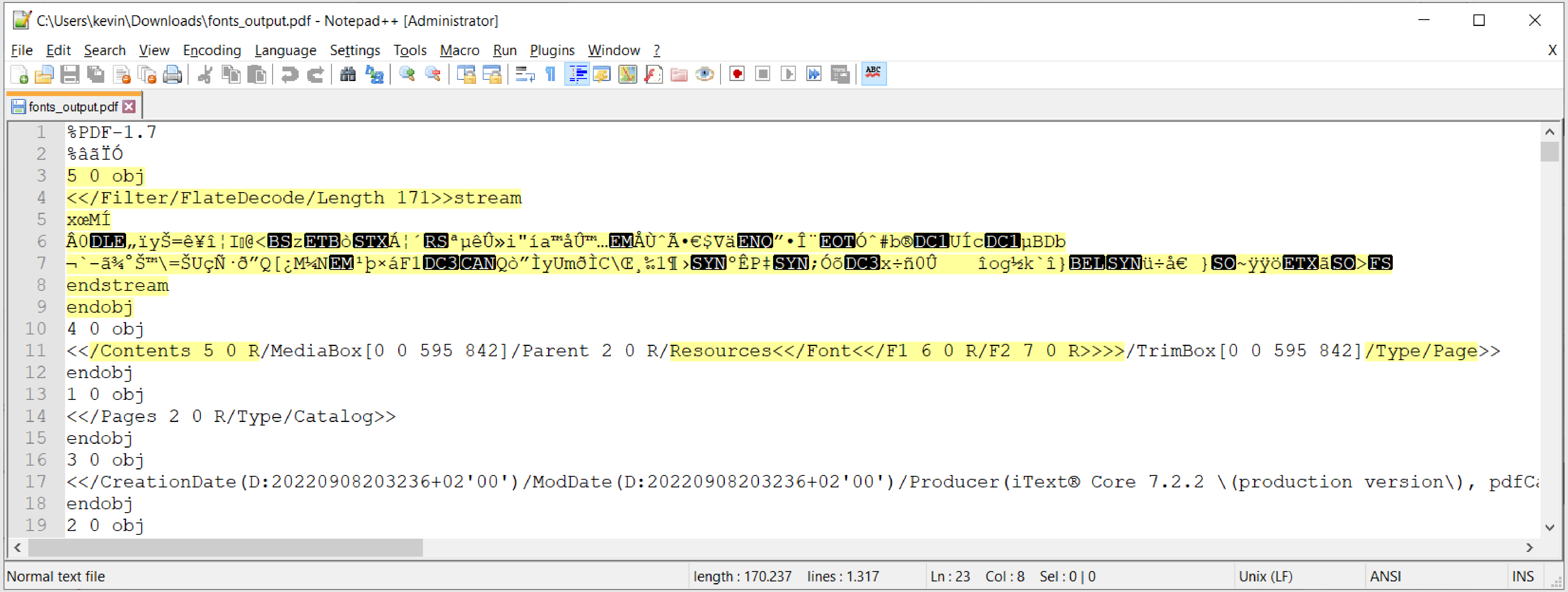

fonts_output.pdf - Adobe Acrobat Pro DC (32-bit)
File Edit View E-Sign Window Help
Home Tools fonts_output.pdf
1 / 1
200%
Find
font
Previous
Next
Replace with
595,0 x 842,0 pt

This text was created with a type 0 font

This text was created with a type 1 font

Case 1 - Debugging

working together!



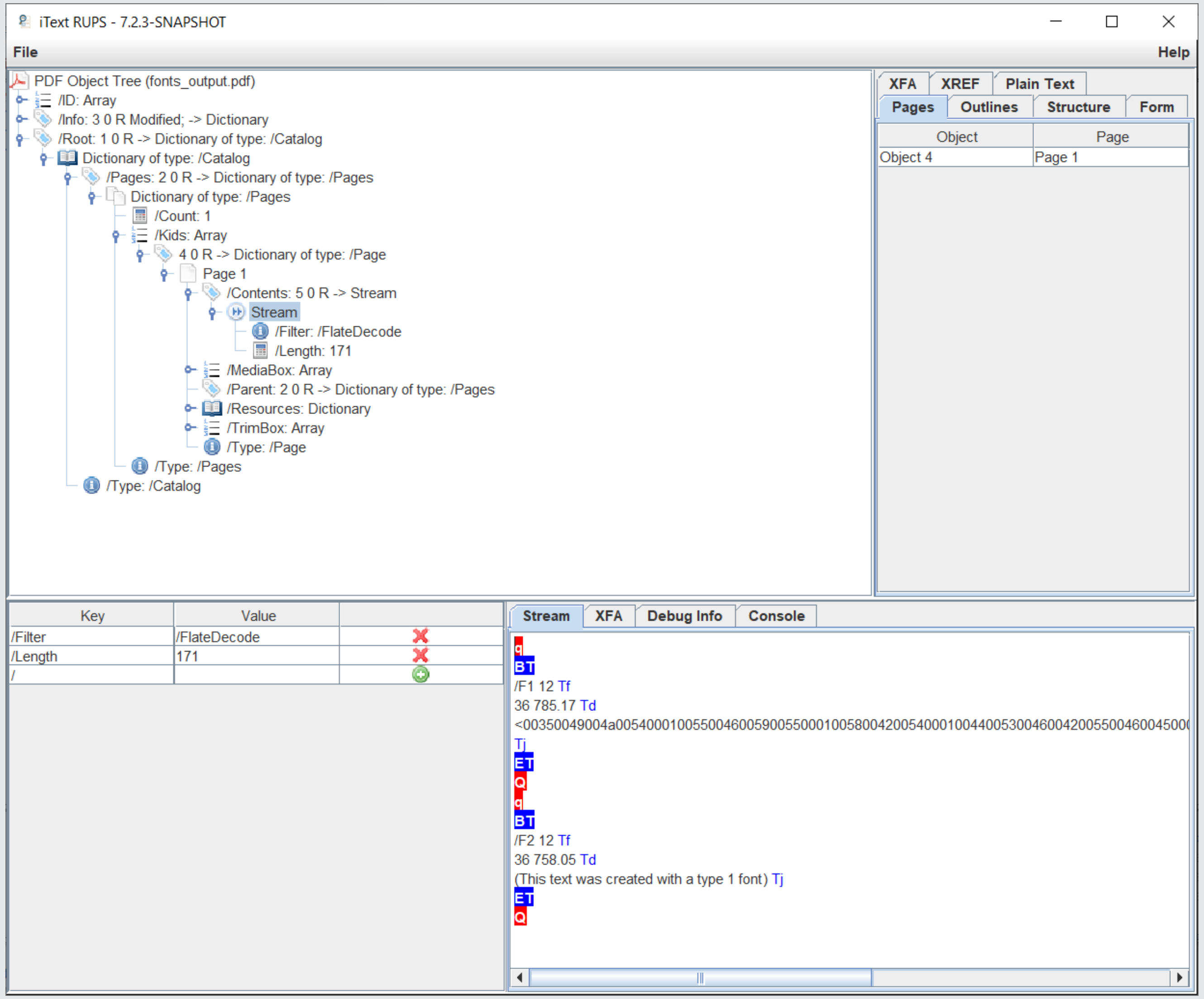
C:\Users\kevin\Downloads\fonts_output.pdf - Notepad++ [Administrator]

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
fonts_output.pdf

```
%PDF-1.7
%âãÏÓ
5 0 obj
<</Filter/FlateDecode/Length 171>>stream
xœMÍ
Â0DLE„ïyŠ=ê¥î¦I▯@<BSzETBòSTXÁ¦´RSªµêÛ»i"ía™åÛ™…EMÅÙ^Ã•€$VäENQ”•Î¨EOTÓ^#b®DC1UÍcDC1µBDb
¬`-ã¾°Š™\=ŠUçÑ·ð”Q[¿M¼NEM¹þ×áF1DC3CANQò”ÌyUmðÌC\Œ¸‰1¶›SYN°ÊP‡SYN;ÓõDC3x÷ñ0Û     îog½k`î}BELSYNü÷å€ }SO~ÿÿöETXãSO>FS
endstream
endobj
4 0 obj
<</Contents 5 0 R/MediaBox[0 0 595 842]/Parent 2 0 R/Resources<</Font<</F1 6 0 R/F2 7 0 R>>>>/TrimBox[0 0 595 842]/Type/Page>>
endobj
1 0 obj
<</Pages 2 0 R/Type/Catalog>>
endobj
3 0 obj
<</CreationDate(D:20220908203236+02'00')/ModDate(D:20220908203236+02'00')/Producer(iText® Core 7.2.2 \(production version\), pdfC
endobj
2 0 obj
```

Normal text file | length : 170.237 lines : 1.317 | Ln : 23 Col : 8 Sel : 0 | 0 | Unix (LF) | ANSI | INS

Case 1 - Debugging

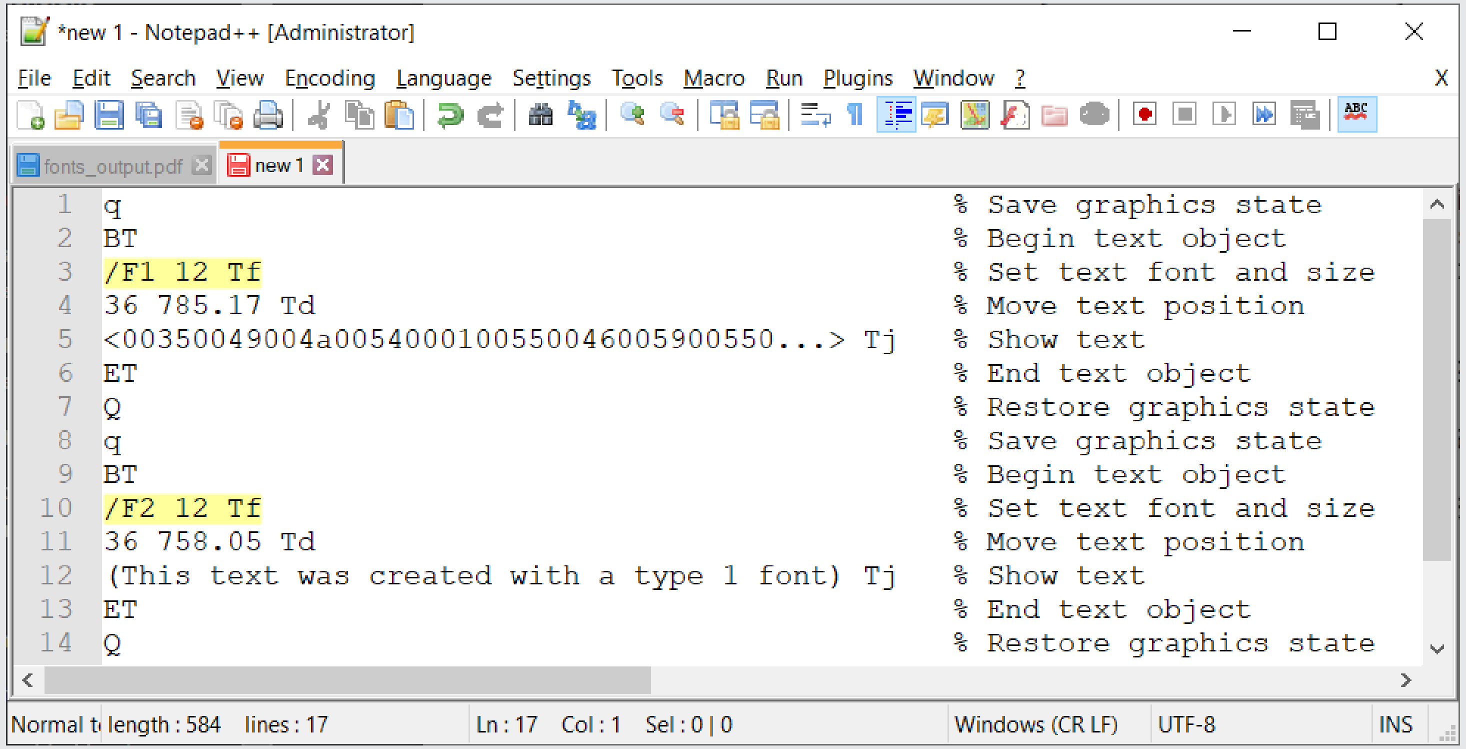


```
PDF Object Tree (fonts_output.pdf)
  /ID: Array
  /Info: 3 0 R Modified; -> Dictionary
  /Root: 1 0 R -> Dictionary of type: /Catalog
    Dictionary of type: /Catalog
      /Pages: 2 0 R -> Dictionary of type: /Pages
        Dictionary of type: /Pages
          /Count: 1
          /Kids: Array
            4 0 R -> Dictionary of type: /Page
              Page 1
                /Contents: 5 0 R -> Stream
                  Stream
                    /Filter: /FlateDecode
                    /Length: 171
                /MediaBox: Array
                /Parent: 2 0 R -> Dictionary of type: /Pages
                /Resources: Dictionary
                /TrimBox: Array
                /Type: /Page
          /Type: /Pages
      /Type: /Catalog
```

Object	Page
Object 4	Page 1

Key	Value	
/Filter	/FlateDecode	✗
/Length	171	✗
/		⊕

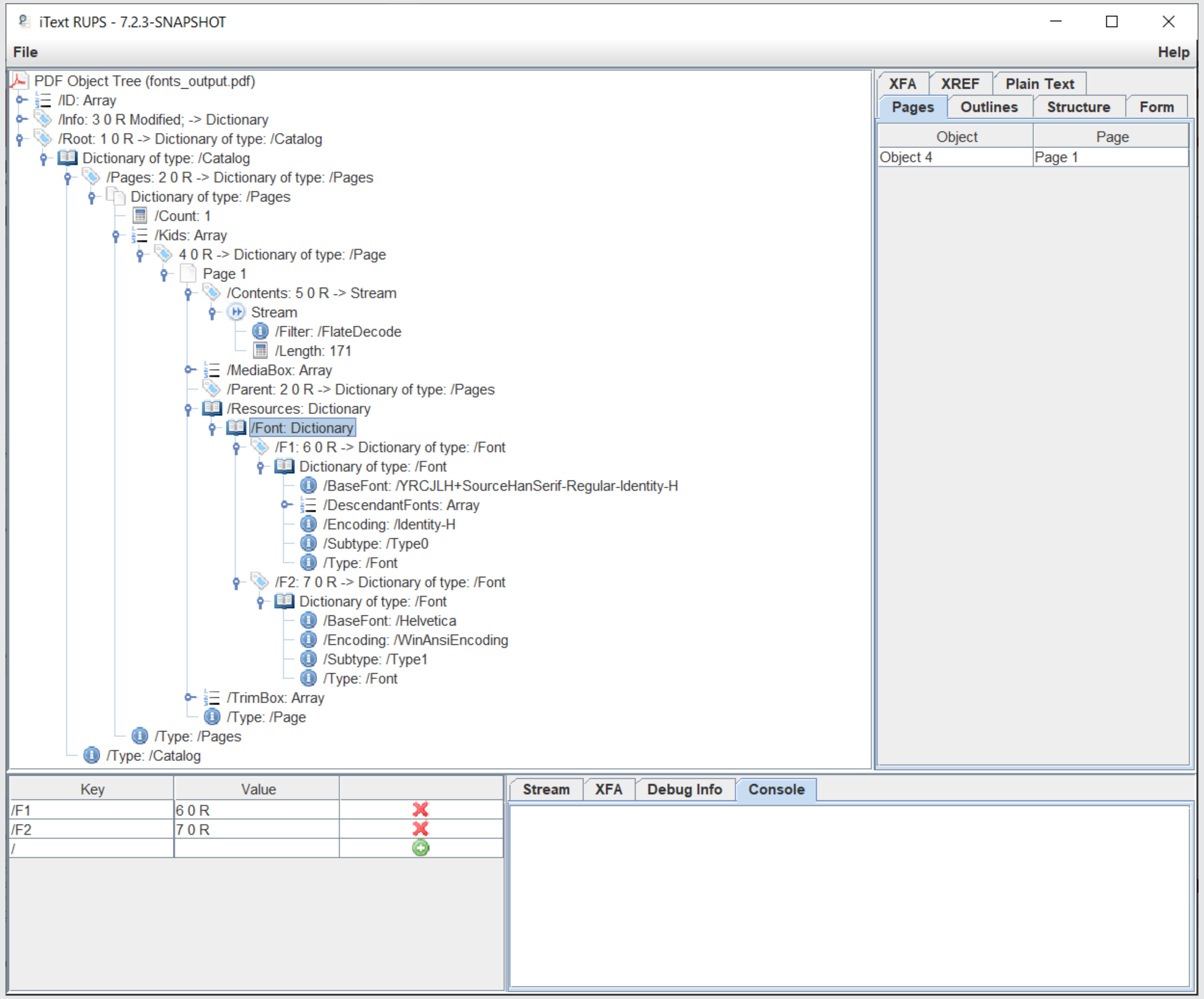
```
q
BT
/F1 12 Tf
36 785.17 Td
<00350049004a00540001005500460059005500010058004200540001004400530046004200550046004500
Tj
ET
Q
q
BT
/F2 12 Tf
36 758.05 Td
(This text was created with a type 1 font) Tj
ET
Q
```



```
q                                                   % Save graphics state
BT                                                  % Begin text object
/F1 12 Tf                                           % Set text font and size
36 785.17 Td                                        % Move text position
<00350049004a005400010055004600590055000...> Tj     % Show text
ET                                                  % End text object
Q                                                   % Restore graphics state
q                                                   % Save graphics state
BT                                                  % Begin text object
/F2 12 Tf                                           % Set text font and size
36 758.05 Td                                        % Move text position
(This text was created with a type 1 font) Tj       % Show text
ET                                                  % End text object
Q                                                   % Restore graphics state
```

Normal t length : 584 lines : 17 Ln : 17 Col : 1 Sel : 0 | 0 Windows (CR LF) UTF-8 INS

Case 1 - Debugging



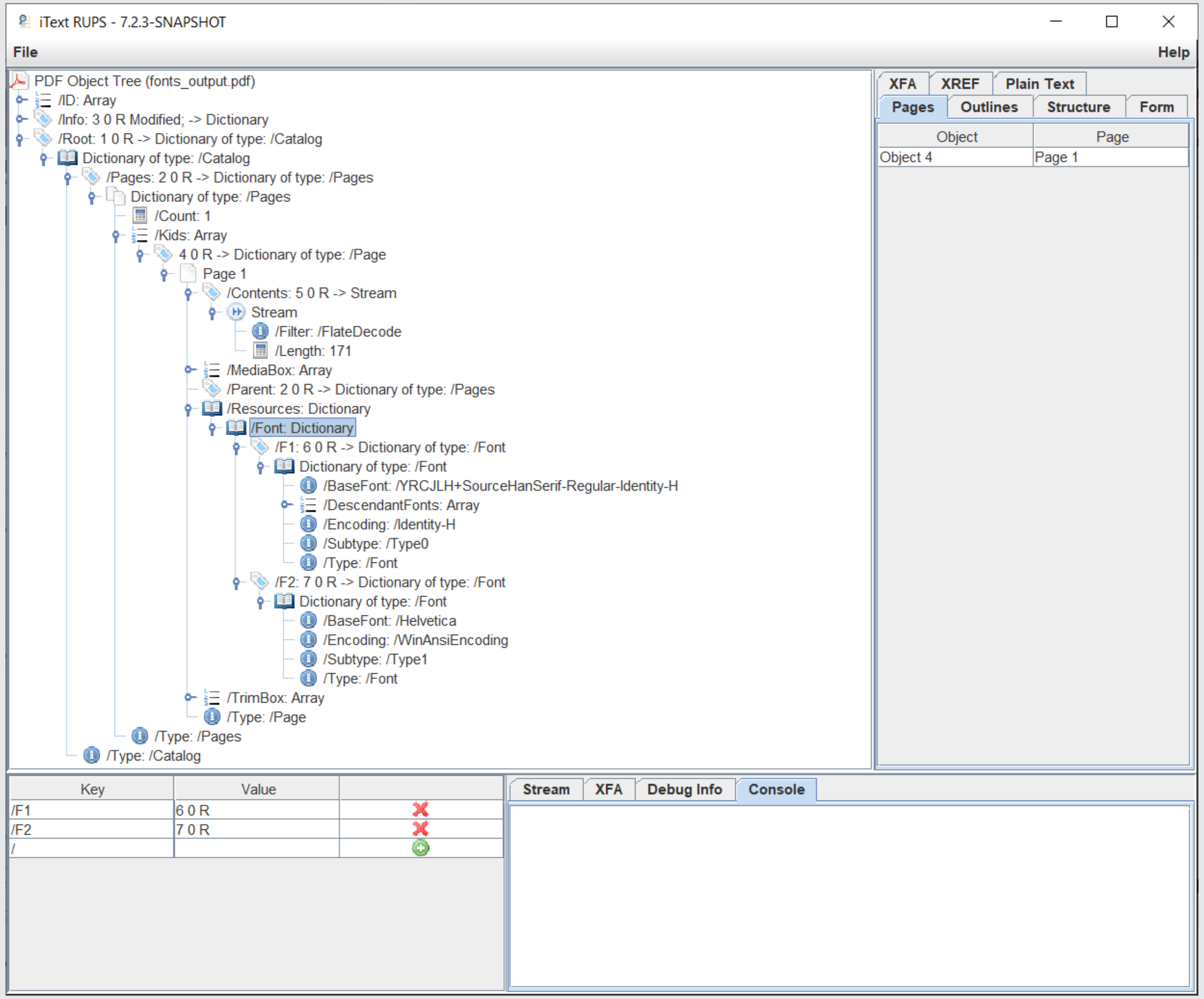
```
PDF Object Tree (fonts_output.pdf)
  /ID: Array
  /Info: 3 0 R Modified; -> Dictionary
  /Root: 1 0 R -> Dictionary of type: /Catalog
    Dictionary of type: /Catalog
      /Pages: 2 0 R -> Dictionary of type: /Pages
        Dictionary of type: /Pages
          /Count: 1
          /Kids: Array
            4 0 R -> Dictionary of type: /Page
              Page 1
                /Contents: 5 0 R -> Stream
                  Stream
                    /Filter: /FlateDecode
                    /Length: 171
                /MediaBox: Array
                /Parent: 2 0 R -> Dictionary of type: /Pages
                /Resources: Dictionary
                  /Font: Dictionary
                    /F1: 6 0 R -> Dictionary of type: /Font
                      Dictionary of type: /Font
                        /BaseFont: /YRCJLH+SourceHanSerif-Regular-Identity-H
                        /DescendantFonts: Array
                        /Encoding: /Identity-H
                        /Subtype: /Type0
                        /Type: /Font
                    /F2: 7 0 R -> Dictionary of type: /Font
                      Dictionary of type: /Font
                        /BaseFont: /Helvetica
                        /Encoding: /WinAnsiEncoding
                        /Subtype: /Type1
                        /Type: /Font
                /TrimBox: Array
                /Type: /Page
          /Type: /Pages
      /Type: /Catalog
```

Object	Page
Object 4	Page 1

Key	Value	
/F1	6 0 R	✗
/F2	7 0 R	✗
/		⊕

Case 1 - Debugging

A **type 0 font** is a virtual font that references multiple component fonts.



```
PDF Object Tree (fonts_output.pdf)
  /ID: Array
  /Info: 3 0 R Modified; -> Dictionary
  /Root: 1 0 R -> Dictionary of type: /Catalog
    Dictionary of type: /Catalog
      /Pages: 2 0 R -> Dictionary of type: /Pages
        Dictionary of type: /Pages
          /Count: 1
          /Kids: Array
            4 0 R -> Dictionary of type: /Page
              Page 1
                /Contents: 5 0 R -> Stream
                  Stream
                    /Filter: /FlateDecode
                    /Length: 171
                /MediaBox: Array
                /Parent: 2 0 R -> Dictionary of type: /Pages
                /Resources: Dictionary
                  /Font: Dictionary
                    /F1: 6 0 R -> Dictionary of type: /Font
                      Dictionary of type: /Font
                        /BaseFont: /YRCJLH+SourceHanSerif-Regular-Identity-H
                        /DescendantFonts: Array
                        /Encoding: /Identity-H
                        /Subtype: /Type0
                        /Type: /Font
                    /F2: 7 0 R -> Dictionary of type: /Font
                      Dictionary of type: /Font
                        /BaseFont: /Helvetica
                        /Encoding: /WinAnsiEncoding
                        /Subtype: /Type1
                        /Type: /Font
                /TrimBox: Array
                /Type: /Page
          /Type: /Pages
      /Type: /Catalog
```

Object	Page
Object 4	Page 1

Key	Value	
/F1	6 0 R	☒
/F2	7 0 R	☒
/		⊕

Case 1 - Debugging

A **type 0 font** is a virtual font that references multiple component fonts.

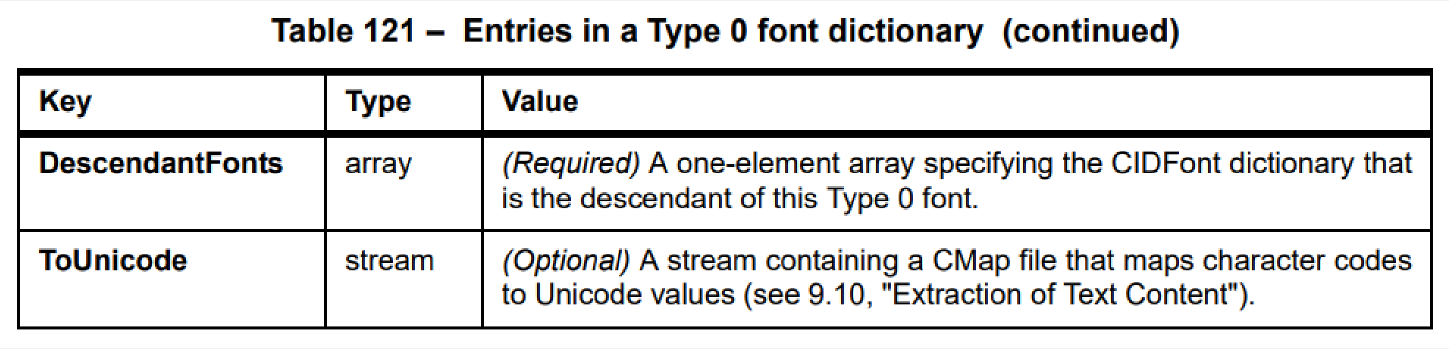
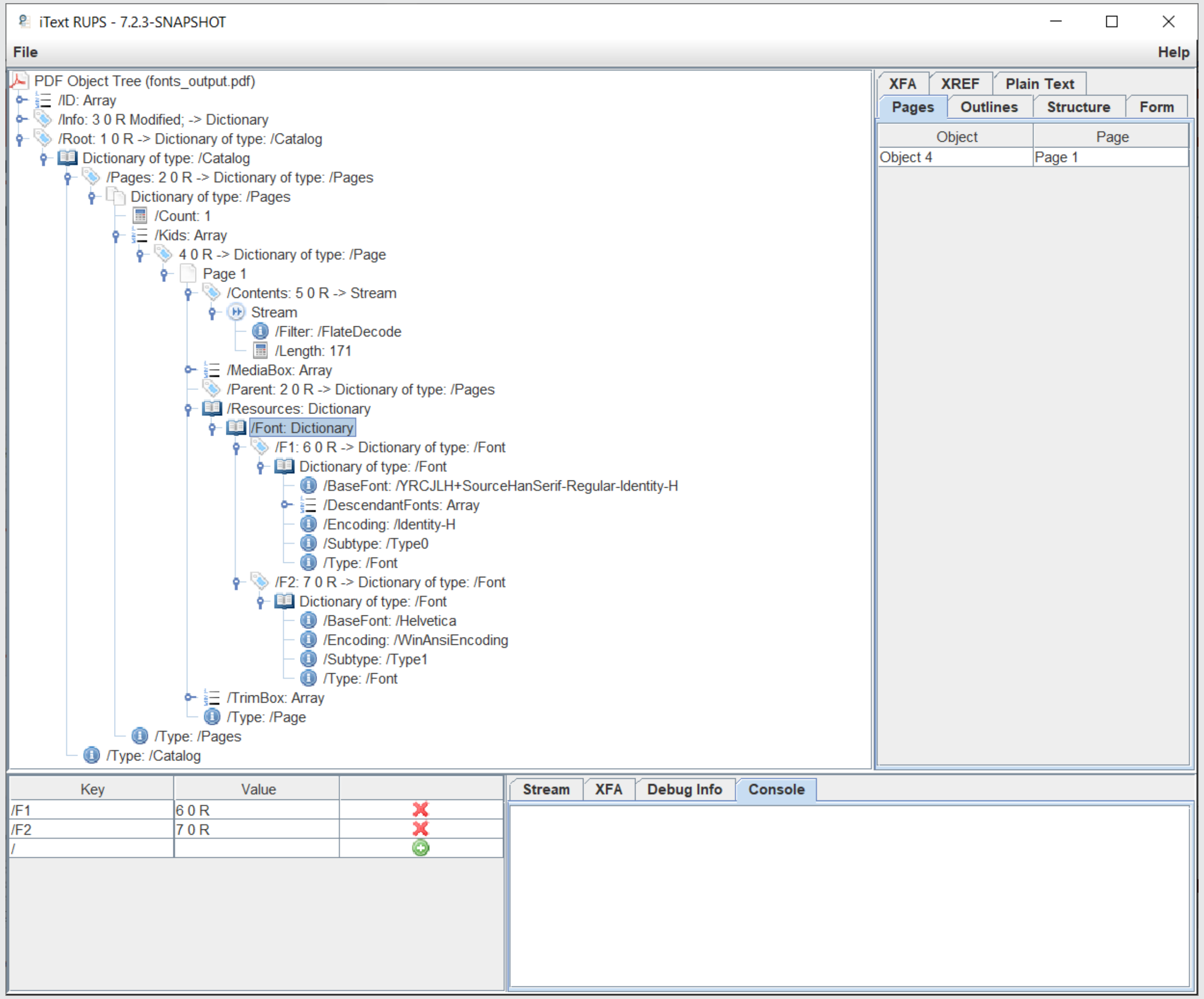


Table 121 – Entries in a Type 0 font dictionary (continued)

Key	Type	Value
DescendantFonts	array	*(Required)* A one-element array specifying the CIDFont dictionary that is the descendant of this Type 0 font.
ToUnicode	stream	*(Optional)* A stream containing a CMap file that maps character codes to Unicode values (see 9.10, "Extraction of Text Content").



```
PDF Object Tree (fonts_output.pdf)
  /ID: Array
  /Info: 3 0 R Modified; -> Dictionary
  /Root: 1 0 R -> Dictionary of type: /Catalog
    Dictionary of type: /Catalog
      /Pages: 2 0 R -> Dictionary of type: /Pages
        Dictionary of type: /Pages
          /Count: 1
          /Kids: Array
            4 0 R -> Dictionary of type: /Page
              Page 1
                /Contents: 5 0 R -> Stream
                  Stream
                    /Filter: /FlateDecode
                    /Length: 171
                /MediaBox: Array
                /Parent: 2 0 R -> Dictionary of type: /Pages
                /Resources: Dictionary
                  /Font: Dictionary
                    /F1: 6 0 R -> Dictionary of type: /Font
                      Dictionary of type: /Font
                        /BaseFont: /YRCJLH+SourceHanSerif-Regular-Identity-H
                        /DescendantFonts: Array
                        /Encoding: /Identity-H
                        /Subtype: /Type0
                        /Type: /Font
                    /F2: 7 0 R -> Dictionary of type: /Font
                      Dictionary of type: /Font
                        /BaseFont: /Helvetica
                        /Encoding: /WinAnsiEncoding
                        /Subtype: /Type1
                        /Type: /Font
                /TrimBox: Array
                /Type: /Page
          /Type: /Pages
      /Type: /Catalog
```

Object	Page
Object 4	Page 1

Key	Value	
/F1	6 0 R	✗
/F2	7 0 R	✗
/		⊕

Case 1 - Debugging

CID fonts are special fonts designed to display languages with more than 256 characters.

The main features that CID fonts add are the ability to have 16bit values (so 65535 separate CID characters rather than 256) and much more sophisticated and more flexible unicode settings for extraction. **Predefined CMAPs (or custom ones embedded by the user) allow for text extraction to provide appropriate values.**

Case 1 - Debugging

A **type0 font** is a virtual font that references multiple component fonts.

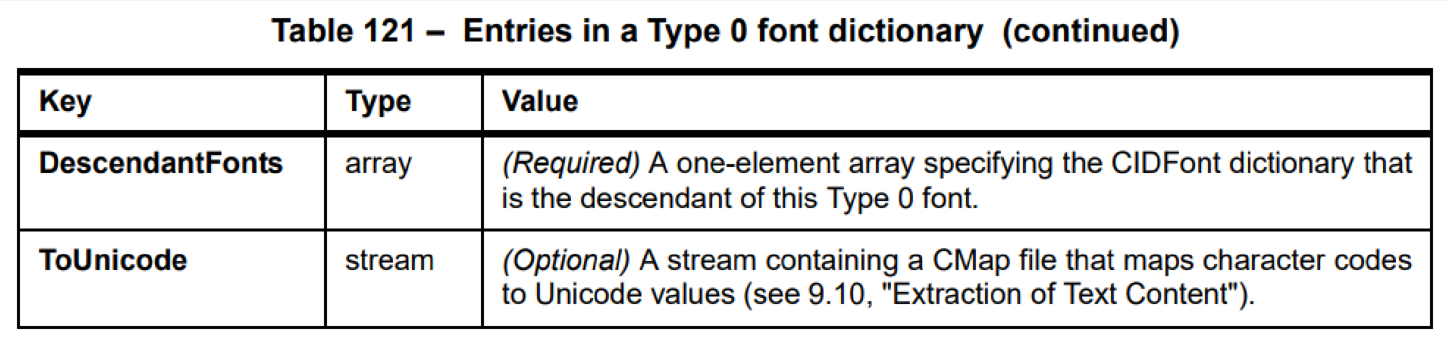
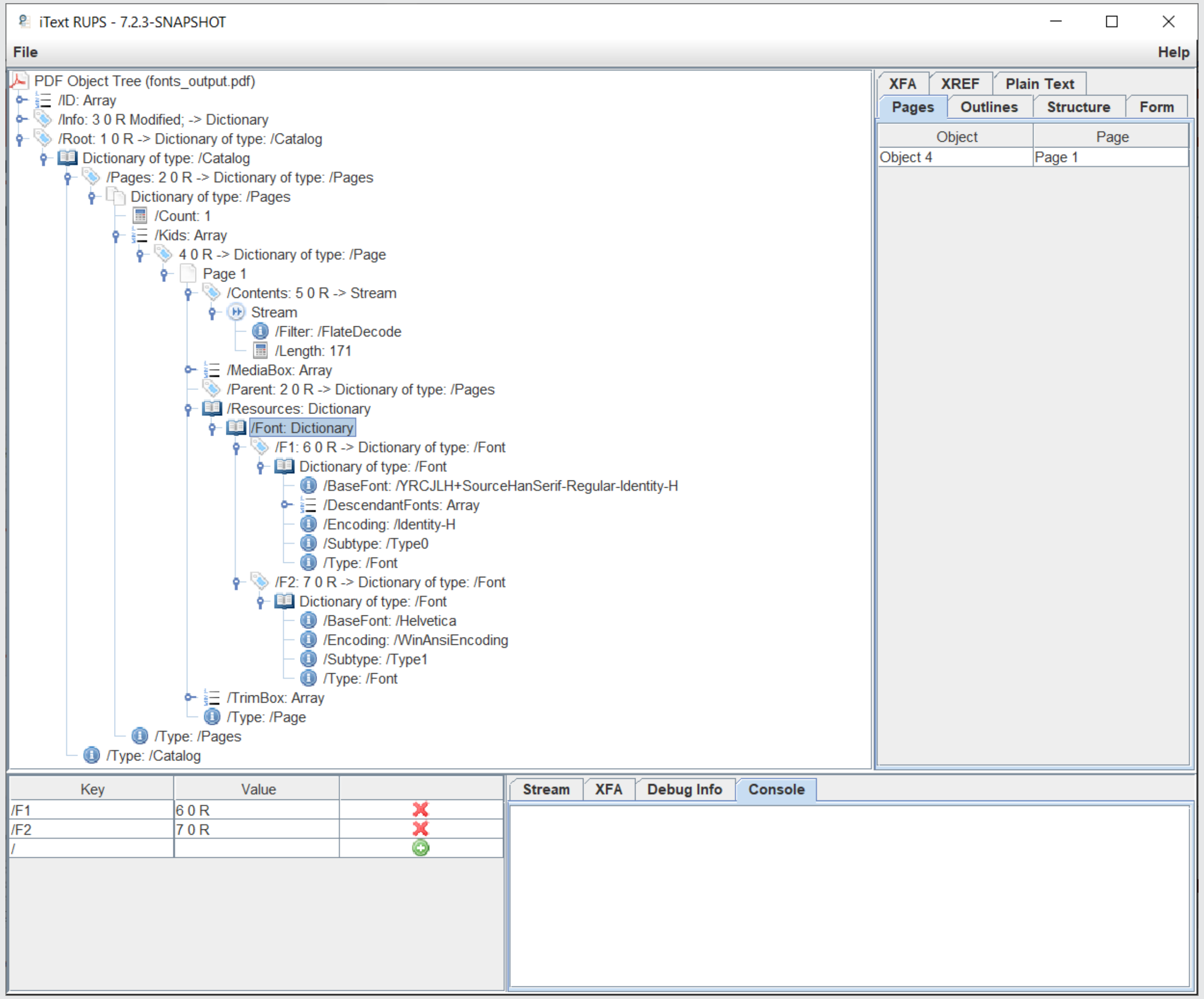


Table 121 – Entries in a Type 0 font dictionary (continued)

Key	Type	Value
DescendantFonts	array	*(Required)* A one-element array specifying the CIDFont dictionary that is the descendant of this Type 0 font.
ToUnicode	stream	*(Optional)* A stream containing a CMap file that maps character codes to Unicode values (see 9.10, "Extraction of Text Content").

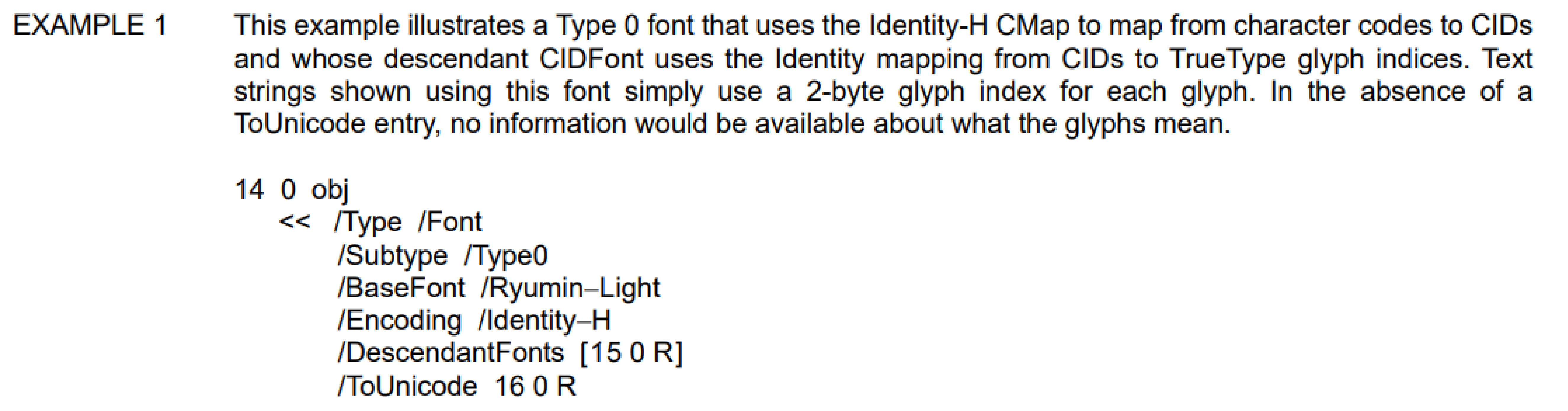


```
PDF Object Tree (fonts_output.pdf)
  /ID: Array
  /Info: 3 0 R Modified; -> Dictionary
  /Root: 1 0 R -> Dictionary of type: /Catalog
    Dictionary of type: /Catalog
      /Pages: 2 0 R -> Dictionary of type: /Pages
        Dictionary of type: /Pages
          /Count: 1
          /Kids: Array
            4 0 R -> Dictionary of type: /Page
              Page 1
                /Contents: 5 0 R -> Stream
                  Stream
                    /Filter: /FlateDecode
                    /Length: 171
                /MediaBox: Array
                /Parent: 2 0 R -> Dictionary of type: /Pages
                /Resources: Dictionary
                  /Font: Dictionary
                    /F1: 6 0 R -> Dictionary of type: /Font
                      Dictionary of type: /Font
                        /BaseFont: /YRCJLH+SourceHanSerif-Regular-Identity-H
                        /DescendantFonts: Array
                        /Encoding: /Identity-H
                        /Subtype: /Type0
                        /Type: /Font
                    /F2: 7 0 R -> Dictionary of type: /Font
                      Dictionary of type: /Font
                        /BaseFont: /Helvetica
                        /Encoding: /WinAnsiEncoding
                        /Subtype: /Type1
                        /Type: /Font
                /TrimBox: Array
                /Type: /Page
          /Type: /Pages
      /Type: /Catalog
```

Object	Page
Object 4	Page 1

Key	Value	
/F1	6 0 R	✗
/F2	7 0 R	✗
/		⊕

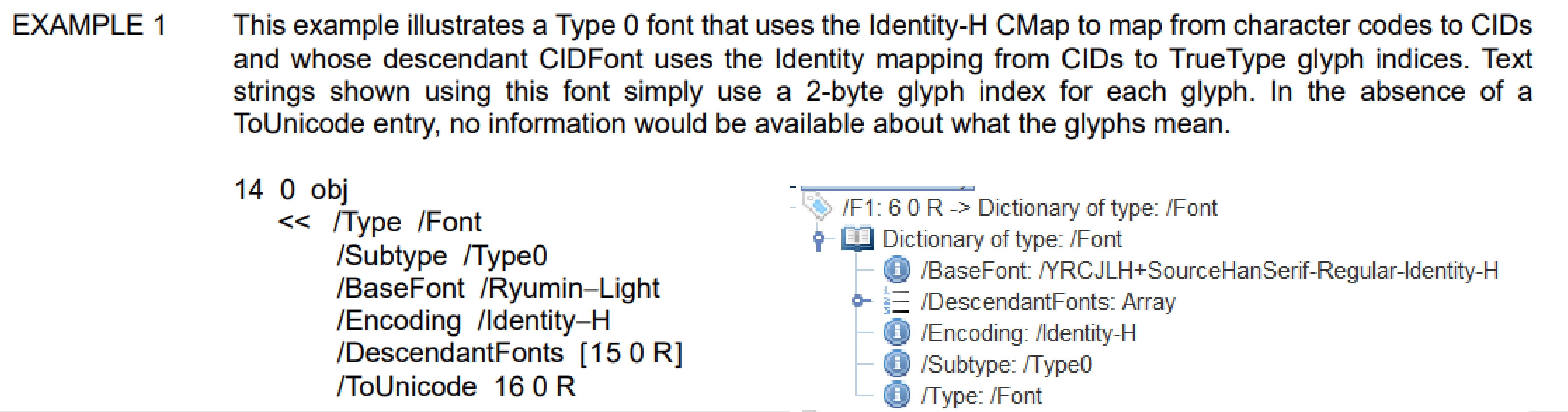
Case 1 - Debugging



EXAMPLE 1 This example illustrates a Type 0 font that uses the Identity-H CMap to map from character codes to CIDs and whose descendant CIDFont uses the Identity mapping from CIDs to TrueType glyph indices. Text strings shown using this font simply use a 2-byte glyph index for each glyph. In the absence of a ToUnicode entry, no information would be available about what the glyphs mean.

```
14  0  obj
    <<  /Type  /Font
        /Subtype  /Type0
        /BaseFont  /Ryumin–Light
        /Encoding  /Identity–H
        /DescendantFonts  [15 0 R]
        /ToUnicode  16 0 R
```

Case 1 - Debugging

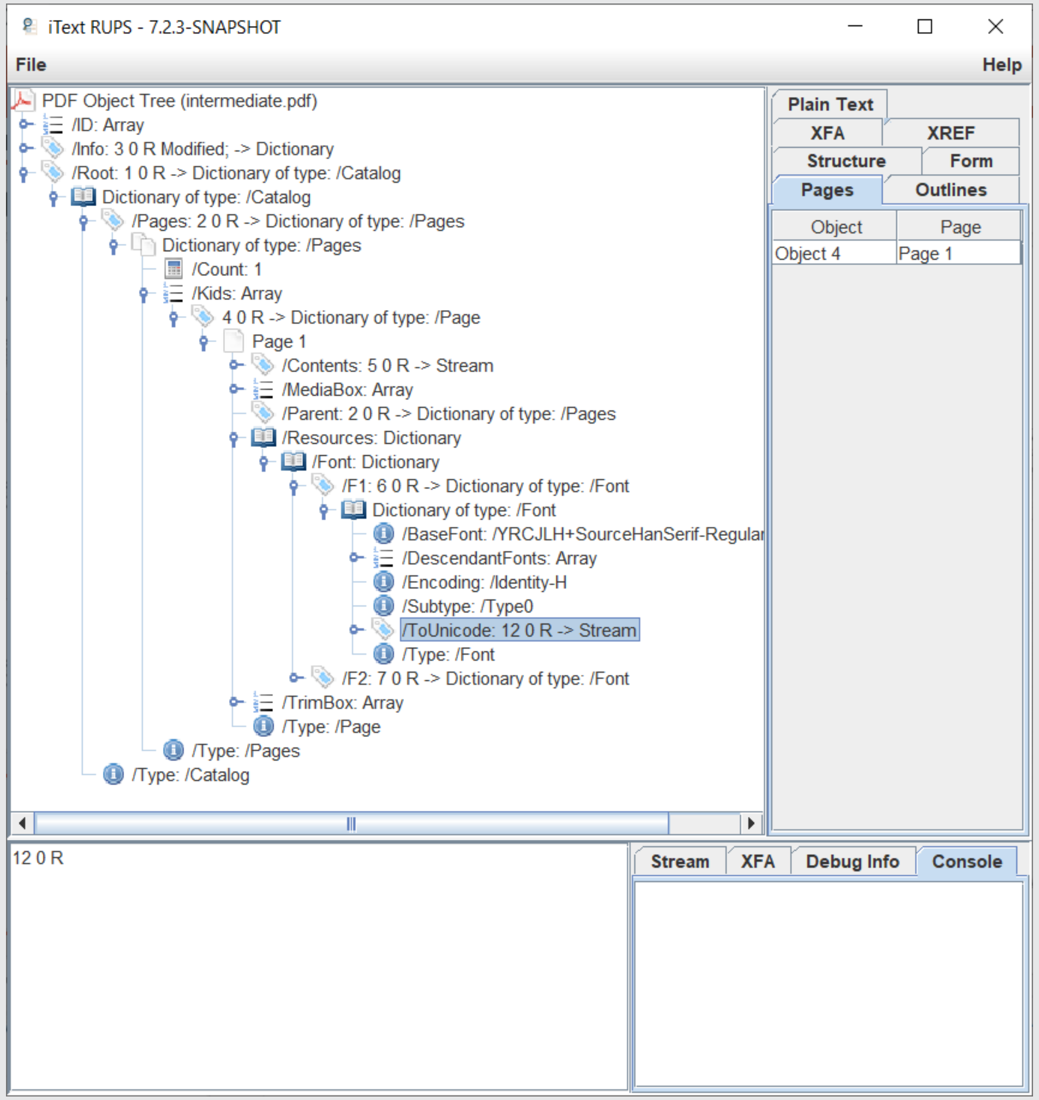


EXAMPLE 1 This example illustrates a Type 0 font that uses the Identity-H CMap to map from character codes to CIDs and whose descendant CIDFont uses the Identity mapping from CIDs to TrueType glyph indices. Text strings shown using this font simply use a 2-byte glyph index for each glyph. In the absence of a ToUnicode entry, no information would be available about what the glyphs mean.

```
14  0  obj
    <<  /Type  /Font
         /Subtype  /Type0
         /BaseFont  /Ryumin–Light
         /Encoding  /Identity–H
         /DescendantFonts  [15 0 R]
         /ToUnicode  16 0 R
```

```
/F1: 6 0 R -> Dictionary of type: /Font
  Dictionary of type: /Font
    /BaseFont: /YRCJLH+SourceHanSerif-Regular-Identity-H
    /DescendantFonts: Array
    /Encoding: /Identity-H
    /Subtype: /Type0
    /Type: /Font
```

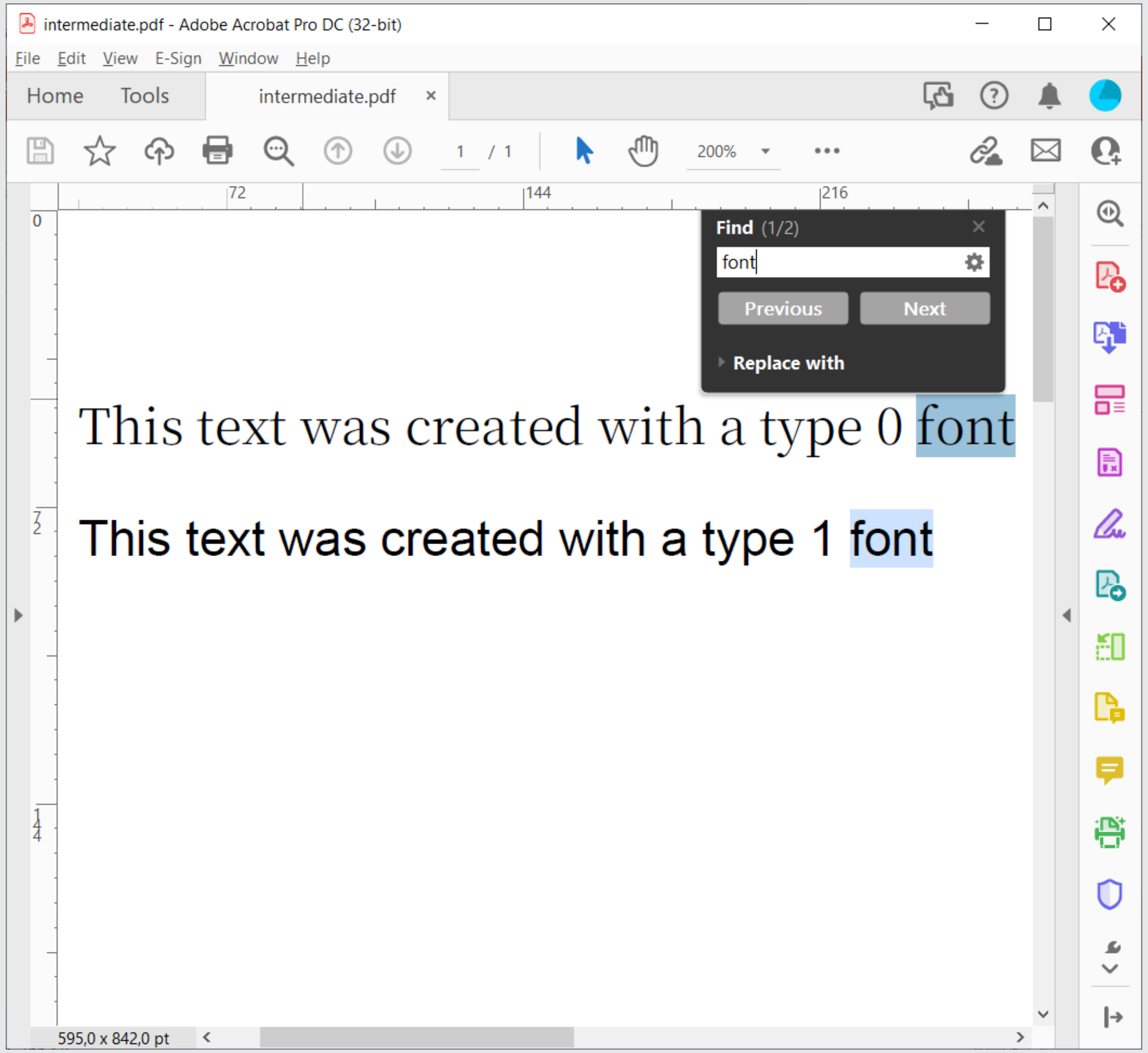
Case 1 - Solution



```
PDF Object Tree (intermediate.pdf)
  /ID: Array
  /Info: 3 0 R Modified; -> Dictionary
  /Root: 1 0 R -> Dictionary of type: /Catalog
    Dictionary of type: /Catalog
      /Pages: 2 0 R -> Dictionary of type: /Pages
        Dictionary of type: /Pages
          /Count: 1
          /Kids: Array
            4 0 R -> Dictionary of type: /Page
              Page 1
                /Contents: 5 0 R -> Stream
                /MediaBox: Array
                /Parent: 2 0 R -> Dictionary of type: /Pages
                /Resources: Dictionary
                  /Font: Dictionary
                    /F1: 6 0 R -> Dictionary of type: /Font
                      Dictionary of type: /Font
                        /BaseFont: /YRCJLH+SourceHanSerif-Regular
                        /DescendantFonts: Array
                        /Encoding: /Identity-H
                        /Subtype: /Type0
                        /ToUnicode: 12 0 R -> Stream
                        /Type: /Font
                    /F2: 7 0 R -> Dictionary of type: /Font
                /TrimBox: Array
                /Type: /Page
          /Type: /Pages
      /Type: /Catalog
```

Object	Page
Object 4	Page 1

12 0 R



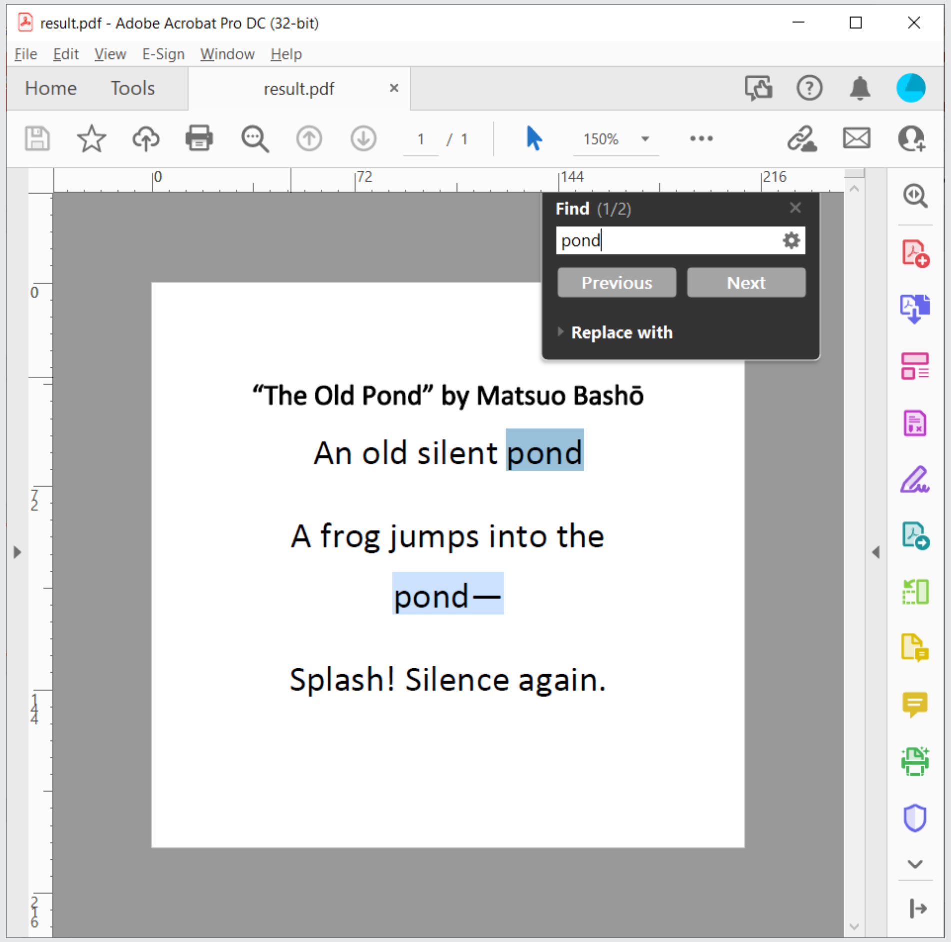
Find: font

This text was created with a type 0 font

This text was created with a type 1 font

Case 2 - Usecase

working together!


result.pdf - Adobe Acrobat Pro DC (32-bit)
File Edit View E-Sign Window Help
Home Tools result.pdf
150%
Find (1/2)
pond
Previous
Next
Replace with

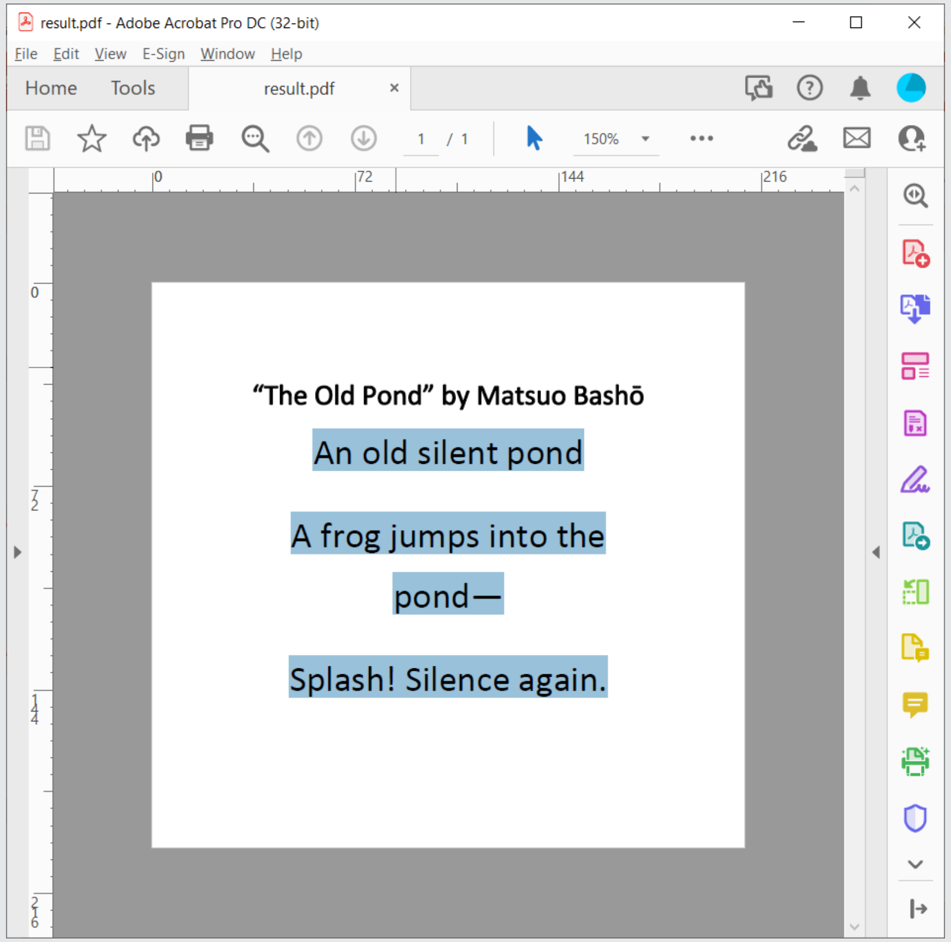
"The Old Pond" by Matsuo Bashō

An old silent pond

A frog jumps into the

pond—

Splash! Silence again.


result.pdf - Adobe Acrobat Pro DC (32-bit)
File Edit View E-Sign Window Help
Home Tools result.pdf
150%

"The Old Pond" by Matsuo Bashō

An old silent pond

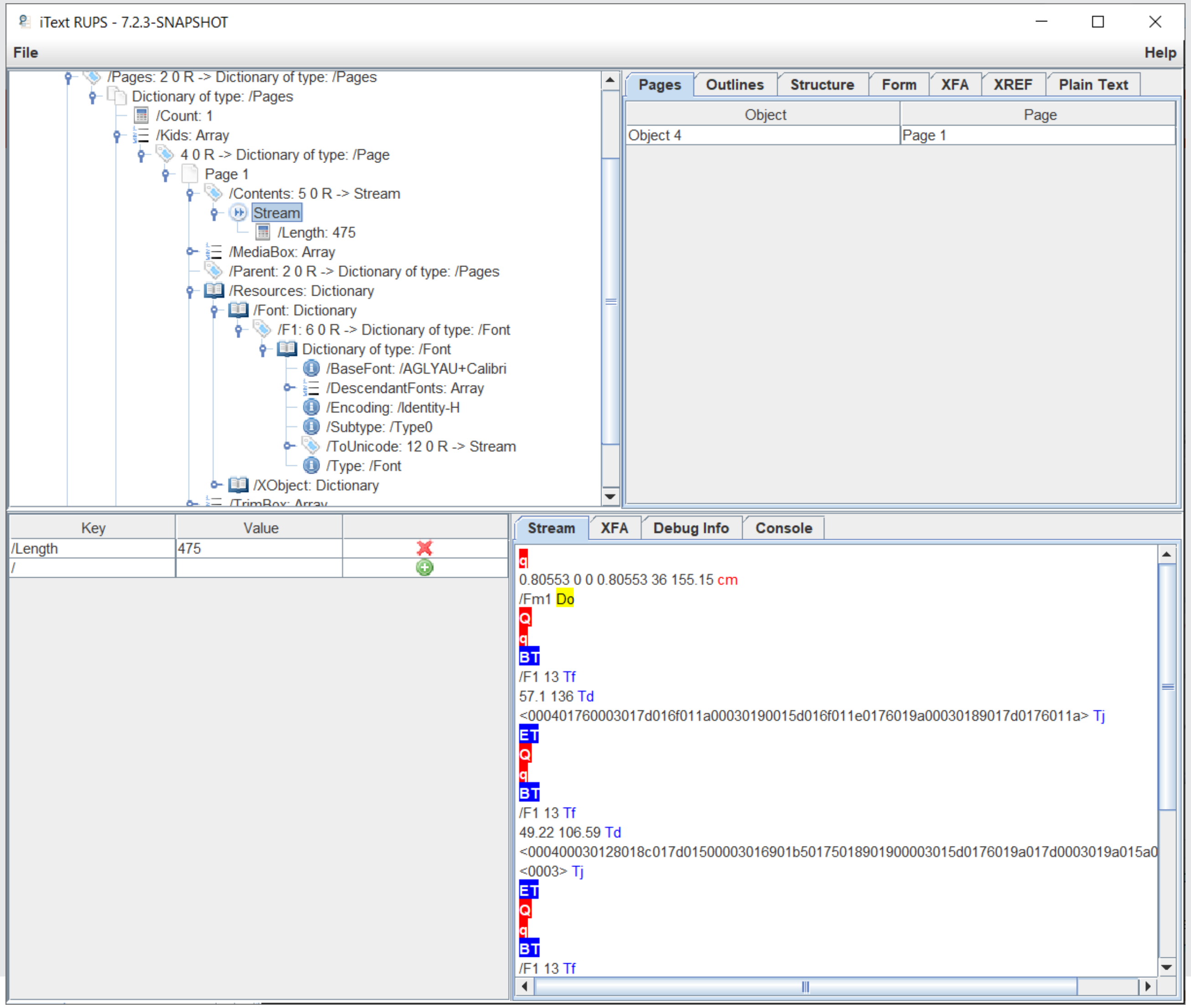
A frog jumps into the

pond—

Splash! Silence again.

PDF association

Case 2 - Debugging



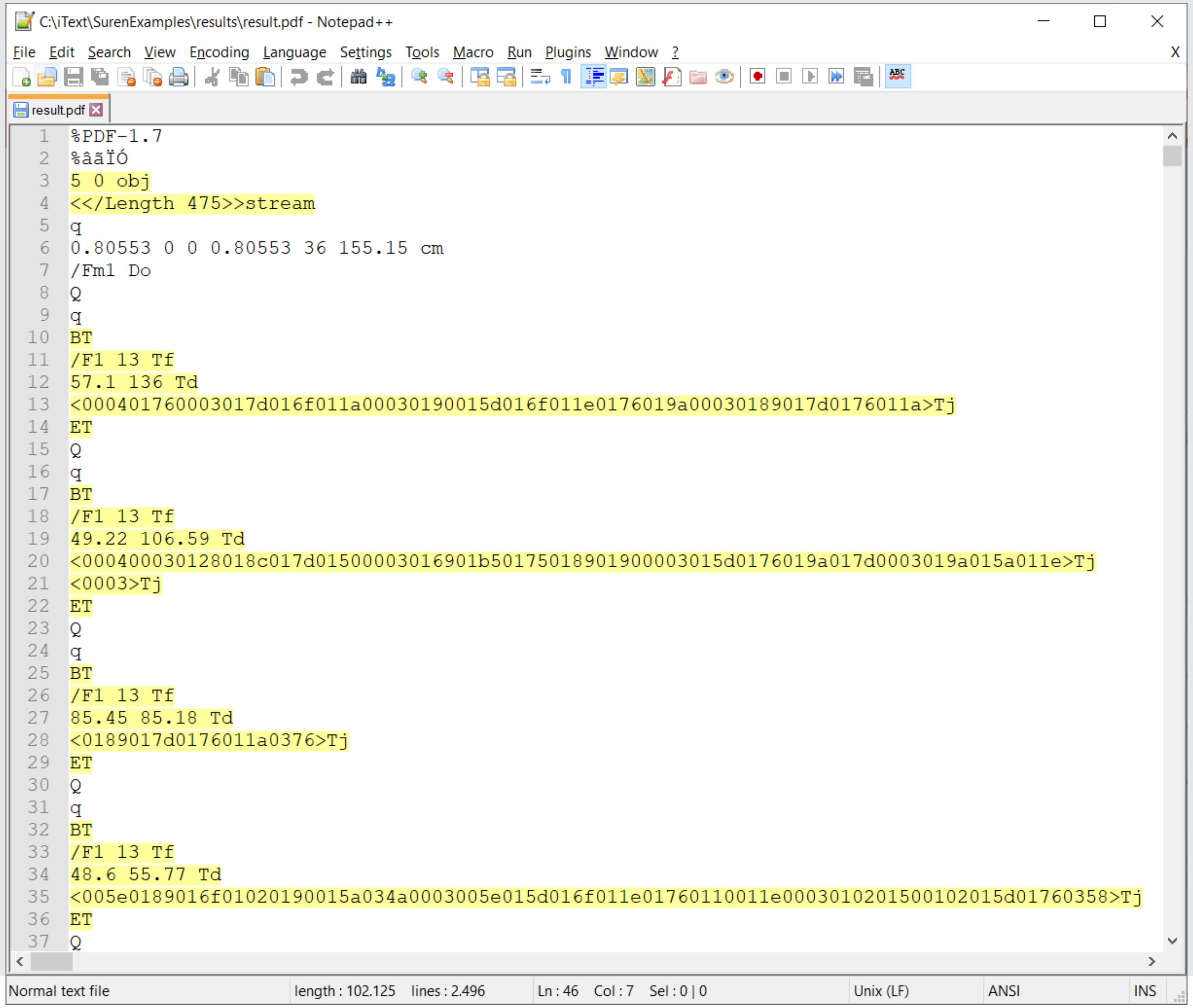
iText RUPS - 7.2.3-SNAPSHOT

```
/Pages: 2 0 R -> Dictionary of type: /Pages
  Dictionary of type: /Pages
    /Count: 1
    /Kids: Array
      4 0 R -> Dictionary of type: /Page
        Page 1
          /Contents: 5 0 R -> Stream
            Stream
              /Length: 475
          /MediaBox: Array
          /Parent: 2 0 R -> Dictionary of type: /Pages
          /Resources: Dictionary
            /Font: Dictionary
              /F1: 6 0 R -> Dictionary of type: /Font
                Dictionary of type: /Font
                  /BaseFont: /AGLYAU+Calibri
                  /DescendantFonts: Array
                  /Encoding: /Identity-H
                  /Subtype: /Type0
                  /ToUnicode: 12 0 R -> Stream
                  /Type: /Font
            /XObject: Dictionary
          /TrimBox: Array
```

Object	Page
Object 4	Page 1

Key	Value	
/Length	475	✗
/		⊕

```
q
0.80553 0 0 0.80553 36 155.15 cm
/Fm1 Do
Q
q
BT
/F1 13 Tf
57.1 136 Td
<000401760003017d016f011a00030190015d016f011e0176019a00030189017d0176011a> Tj
ET
Q
q
BT
/F1 13 Tf
49.22 106.59 Td
<000400030128018c017d01500003016901b50175018901900003015d0176019a017d0003019a015a0
<0003> Tj
ET
Q
q
BT
/F1 13 Tf
```



C:\iText\SurenExamples\results\result.pdf - Notepad++

```
%PDF-1.7
%âãÏÓ
5 0 obj
<</Length 475>>stream
q
0.80553 0 0 0.80553 36 155.15 cm
/Fm1 Do
Q
q
BT
/F1 13 Tf
57.1 136 Td
<000401760003017d016f011a00030190015d016f011e0176019a00030189017d0176011a>Tj
ET
Q
q
BT
/F1 13 Tf
49.22 106.59 Td
<000400030128018c017d01500003016901b50175018901900003015d0176019a017d0003019a015a011e>Tj
<0003>Tj
ET
Q
q
BT
/F1 13 Tf
85.45 85.18 Td
<0189017d0176011a0376>Tj
ET
Q
q
BT
/F1 13 Tf
48.6 55.77 Td
<005e0189016f01020190015a034a0003005e015d016f011e01760110011e000301020150010201 5d01760358>Tj
ET
Q
```

Normal text file | length : 102.125 lines : 2.496 | Ln : 46 Col : 7 Sel : 0 | 0 | Unix (LF) | ANSI | INS

Case 2 - Debugging

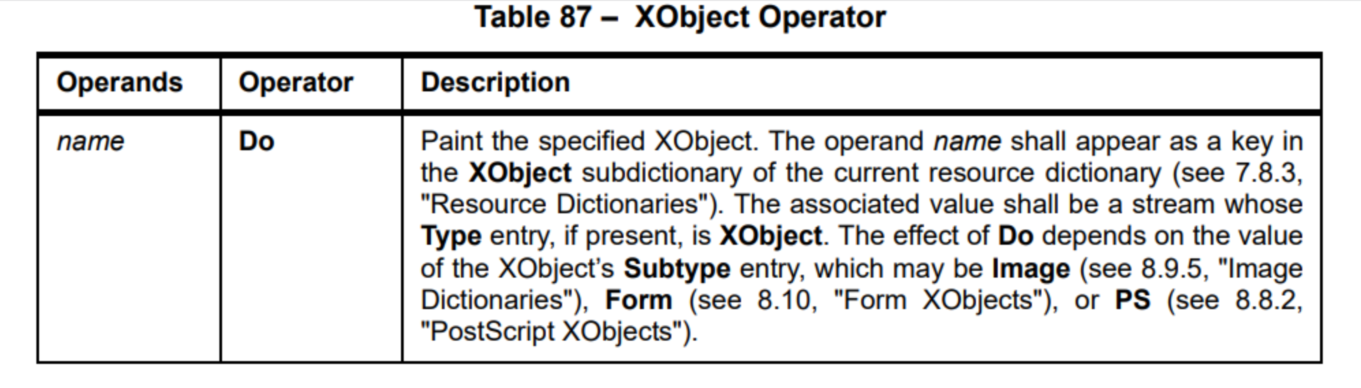
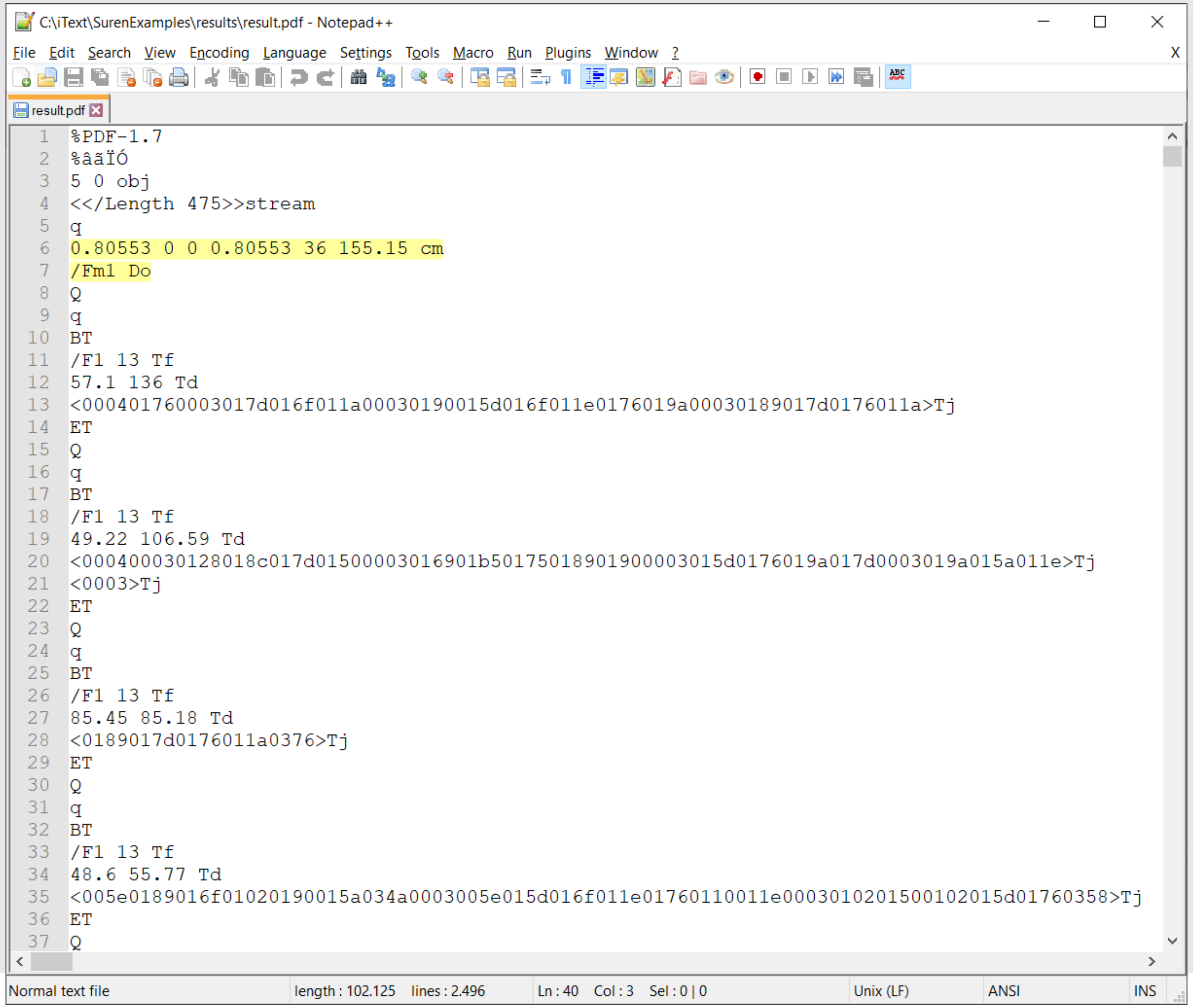


Table 87 – XObject Operator

Operands	Operator	Description
name	**Do**	Paint the specified XObject. The operand *name* shall appear as a key in the **XObject** subdictionary of the current resource dictionary (see 7.8.3, "Resource Dictionaries"). The associated value shall be a stream whose **Type** entry, if present, is **XObject**. The effect of **Do** depends on the value of the XObject's **Subtype** entry, which may be **Image** (see 8.9.5, "Image Dictionaries"), **Form** (see 8.10, "Form XObjects"), or **PS** (see 8.8.2, "PostScript XObjects").

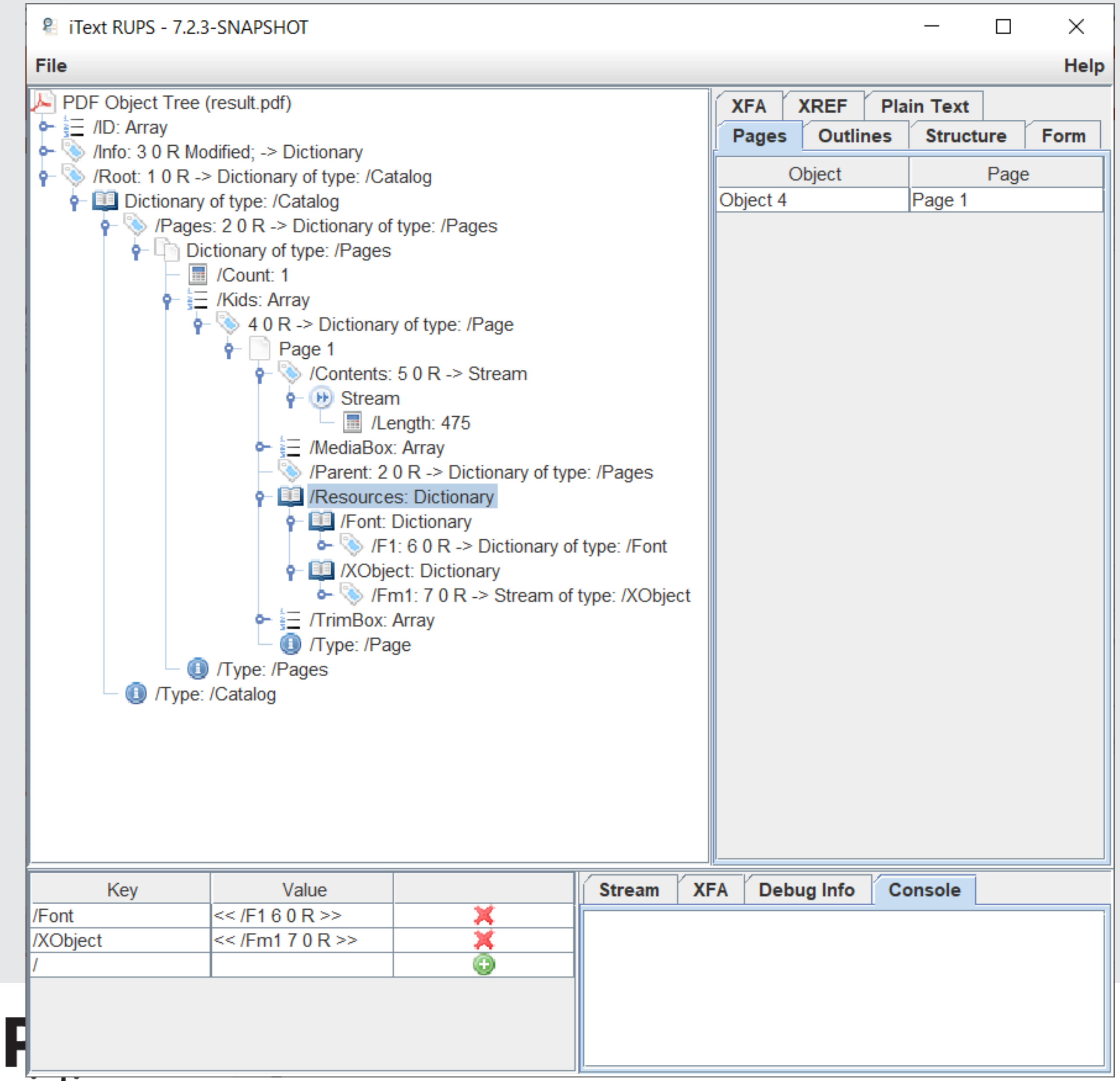


C:\iText\SurenExamples\results\result.pdf - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
result.pdf

```
%PDF-1.7
%âãÏÓ
5 0 obj
<</Length 475>>stream
q
0.80553 0 0 0.80553 36 155.15 cm
/Fm1 Do
Q
q
BT
/F1 13 Tf
57.1 136 Td
<000401760003017d016f011a00030190015d016f011e0176019a00030189017d0176011a>Tj
ET
Q
q
BT
/F1 13 Tf
49.22 106.59 Td
<00040003012801 8c017d01500003016901b50175018901900003015d0176019a017d0003019a015a011e>Tj
<0003>Tj
ET
Q
q
BT
/F1 13 Tf
85.45 85.18 Td
<0189017d0176011a0376>Tj
ET
Q
q
BT
/F1 13 Tf
48.6 55.77 Td
<005e0189016f01020190015a034a0003005e015d016f011e01760110011e00030102015001020 15d01760358>Tj
ET
Q
```

Normal text file   length : 102.125   lines : 2.496   Ln : 40   Col : 3   Sel : 0 | 0   Unix (LF)   ANSI   INS

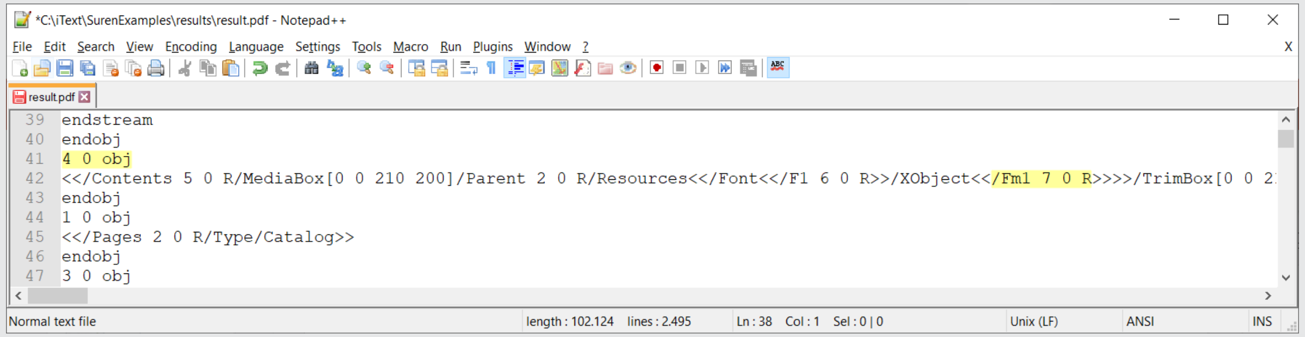
Case 2 - Debugging



```
PDF Object Tree (result.pdf)
  /ID: Array
  /Info: 3 0 R Modified; -> Dictionary
  /Root: 1 0 R -> Dictionary of type: /Catalog
    Dictionary of type: /Catalog
      /Pages: 2 0 R -> Dictionary of type: /Pages
        Dictionary of type: /Pages
          /Count: 1
          /Kids: Array
            4 0 R -> Dictionary of type: /Page
              Page 1
                /Contents: 5 0 R -> Stream
                  Stream
                    /Length: 475
                /MediaBox: Array
                /Parent: 2 0 R -> Dictionary of type: /Pages
                /Resources: Dictionary
                  /Font: Dictionary
                    /F1: 6 0 R -> Dictionary of type: /Font
                  /XObject: Dictionary
                    /Fm1: 7 0 R -> Stream of type: /XObject
                /TrimBox: Array
                /Type: /Page
          /Type: /Pages
      /Type: /Catalog
```

Object	Page
Object 4	Page 1

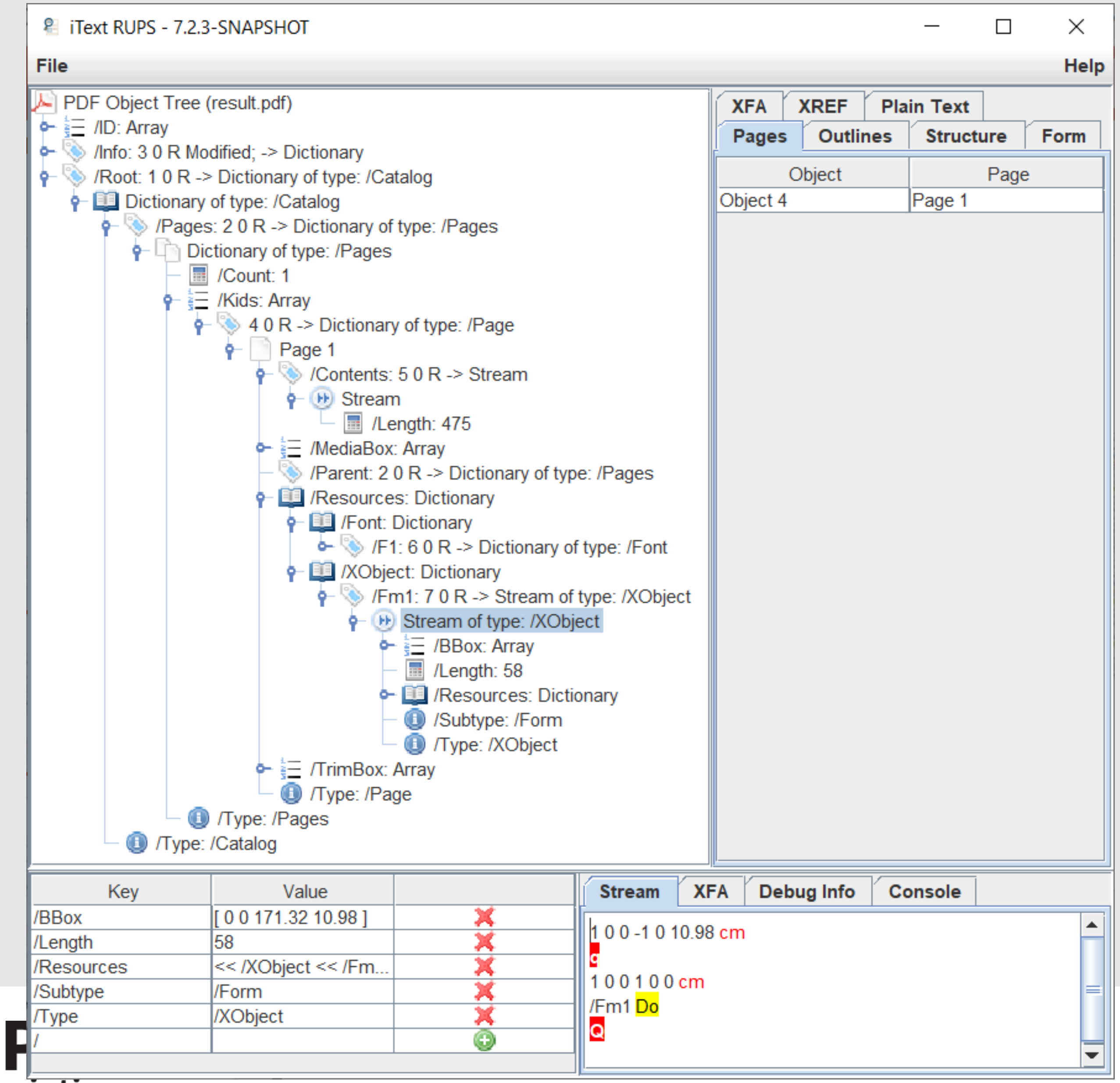
Key	Value	
/Font	<< /F1 6 0 R >>	✗
/XObject	<< /Fm1 7 0 R >>	✗
/		⊕


*C:\iText\SurenExamples\results\result.pdf - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
result.pdf

```
endstream
endobj
4 0 obj
<</Contents 5 0 R/MediaBox[0 0 210 200]/Parent 2 0 R/Resources<</Font<</F1 6 0 R>>/XObject<</Fm1 7 0 R>>>>/TrimBox[0 0 2
endobj
1 0 obj
<</Pages 2 0 R/Type/Catalog>>
endobj
3 0 obj
```

Normal text file | length : 102.124 lines : 2.495 | Ln : 38 Col : 1 Sel : 0 | 0 | Unix (LF) | ANSI | INS

Case 2 - Debugging

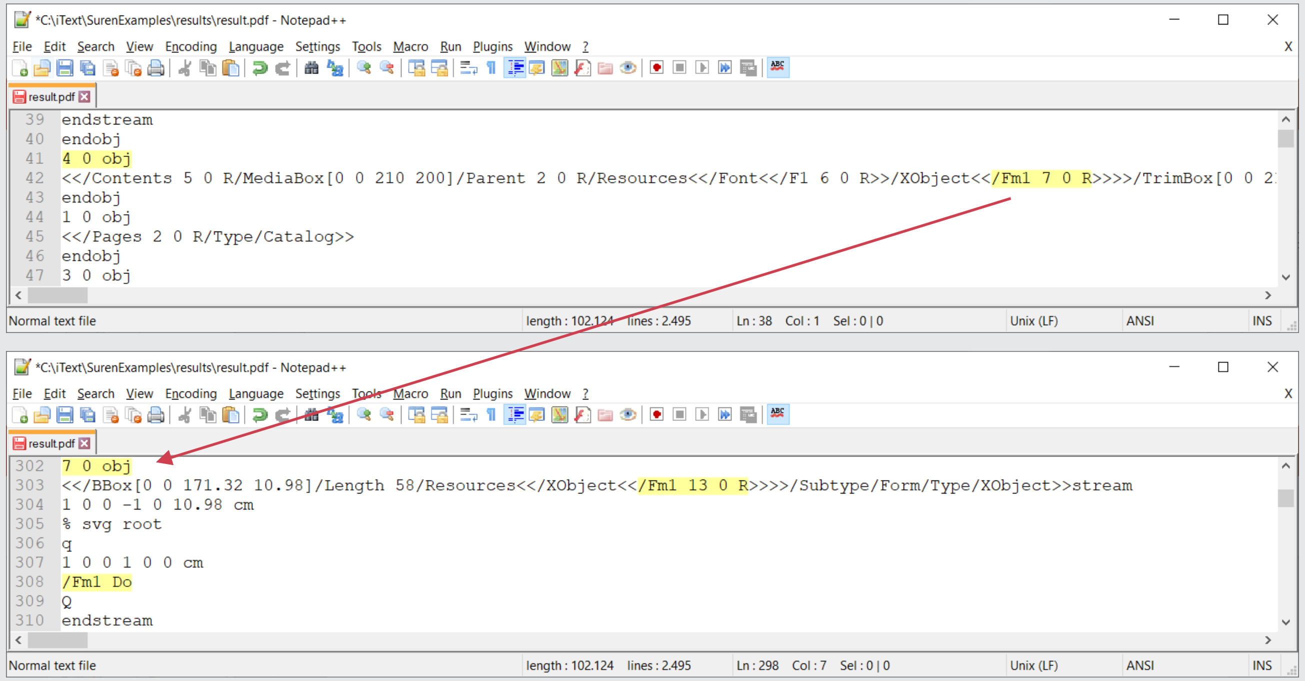


```
PDF Object Tree (result.pdf)
  /ID: Array
  /Info: 3 0 R Modified; -> Dictionary
  /Root: 1 0 R -> Dictionary of type: /Catalog
    Dictionary of type: /Catalog
      /Pages: 2 0 R -> Dictionary of type: /Pages
        Dictionary of type: /Pages
          /Count: 1
          /Kids: Array
            4 0 R -> Dictionary of type: /Page
              Page 1
                /Contents: 5 0 R -> Stream
                  Stream
                    /Length: 475
                /MediaBox: Array
                /Parent: 2 0 R -> Dictionary of type: /Pages
                /Resources: Dictionary
                  /Font: Dictionary
                    /F1: 6 0 R -> Dictionary of type: /Font
                  /XObject: Dictionary
                    /Fm1: 7 0 R -> Stream of type: /XObject
                      Stream of type: /XObject
                        /BBox: Array
                        /Length: 58
                        /Resources: Dictionary
                        /Subtype: /Form
                        /Type: /XObject
                /TrimBox: Array
                /Type: /Page
          /Type: /Pages
      /Type: /Catalog
```

Object	Page
Object 4	Page 1

Key	Value	
/BBox	[0 0 171.32 10.98]	✗
/Length	58	✗
/Resources	<< /XObject << /Fm...	✗
/Subtype	/Form	✗
/Type	/XObject	✗
/		⊕

```
1 0 0 -1 0 10.98 cm
q
1 0 0 1 0 0 cm
/Fm1 Do
Q
```

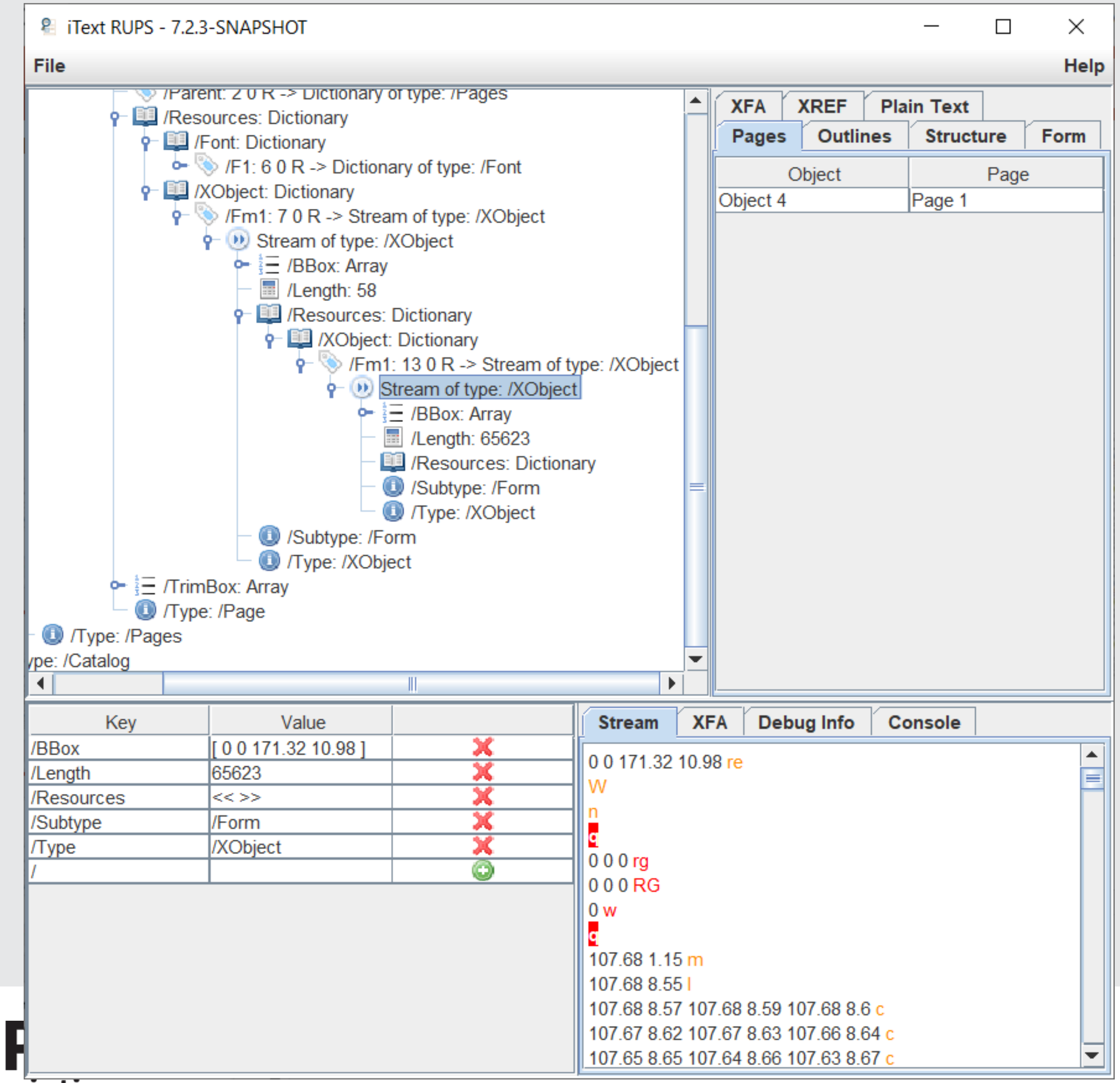

*C:\iText\SurenExamples\results\result.pdf - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
result.pdf
Normal text file
length : 102.124 lines : 2.495
Ln : 38 Col : 1 Sel : 0 | 0
Unix (LF)
ANSI
INS

```
endstream
endobj
4 0 obj
<</Contents 5 0 R/MediaBox[0 0 210 200]/Parent 2 0 R/Resources<</Font<</F1 6 0 R>>/XObject<</Fm1 7 0 R>>>>/TrimBox[0 0 2
endobj
1 0 obj
<</Pages 2 0 R/Type/Catalog>>
endobj
3 0 obj
```

*C:\iText\SurenExamples\results\result.pdf - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
result.pdf
Normal text file
length : 102.124 lines : 2.495
Ln : 298 Col : 7 Sel : 0 | 0
Unix (LF)
ANSI
INS

```
7 0 obj
<</BBox[0 0 171.32 10.98]/Length 58/Resources<</XObject<</Fm1 13 0 R>>>>/Subtype/Form/Type/XObject>>stream
1 0 0 -1 0 10.98 cm
% svg root
q
1 0 0 1 0 0 cm
/Fm1 Do
Q
endstream
```

Case 2 - Debugging

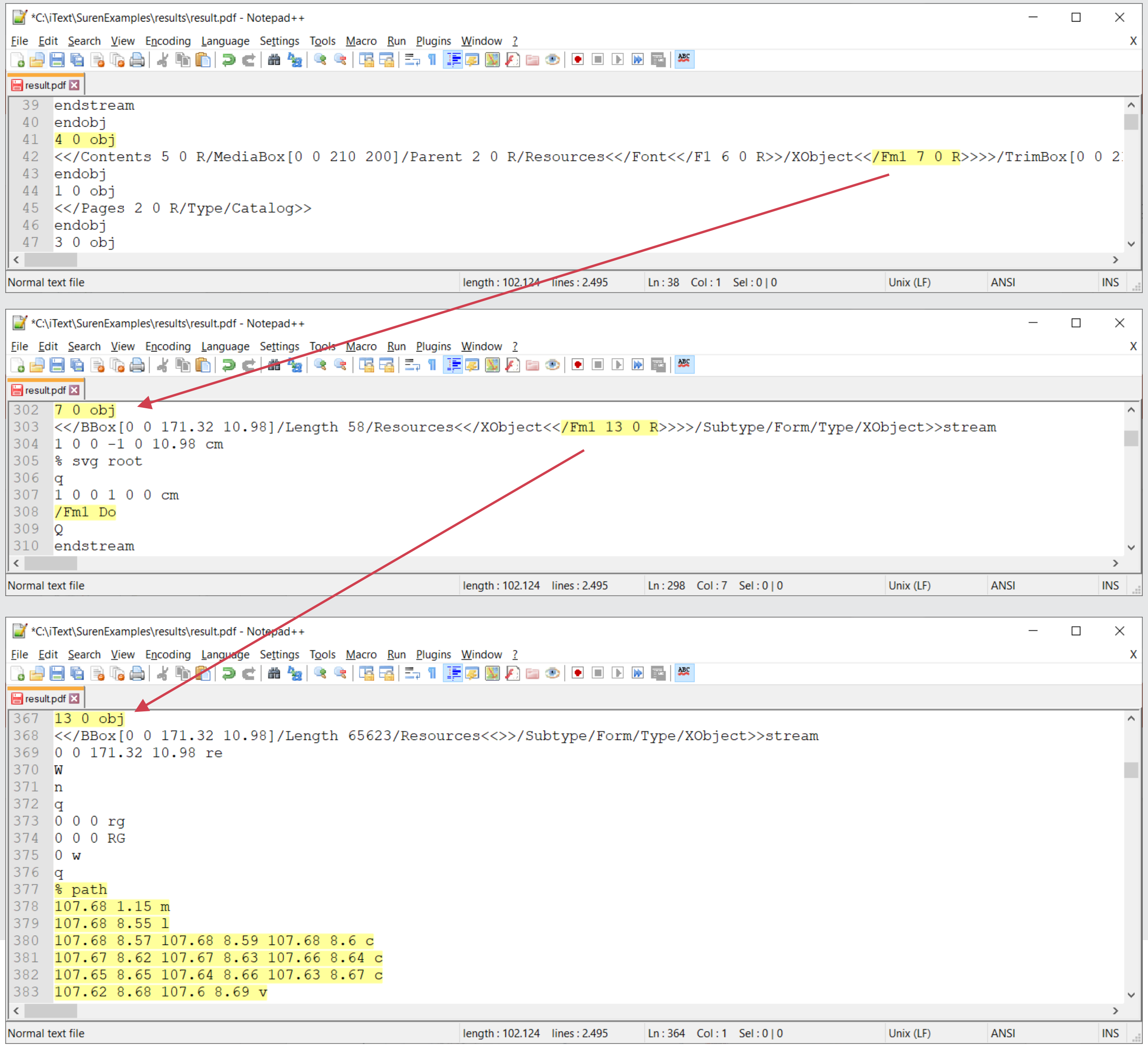


```
/Parent: 2 0 R -> Dictionary of type: /Pages
/Resources: Dictionary
  /Font: Dictionary
    /F1: 6 0 R -> Dictionary of type: /Font
  /XObject: Dictionary
    /Fm1: 7 0 R -> Stream of type: /XObject
      Stream of type: /XObject
        /BBox: Array
        /Length: 58
        /Resources: Dictionary
          /XObject: Dictionary
            /Fm1: 13 0 R -> Stream of type: /XObject
              Stream of type: /XObject
                /BBox: Array
                /Length: 65623
                /Resources: Dictionary
                /Subtype: /Form
                /Type: /XObject
        /Subtype: /Form
        /Type: /XObject
/TrimBox: Array
/Type: /Page
/Type: /Pages
pe: /Catalog
```

Object	Page
Object 4	Page 1

Key	Value	
/BBox	[0 0 171.32 10.98]	☒
/Length	65623	☒
/Resources	<< >>	☒
/Subtype	/Form	☒
/Type	/XObject	☒
/		⊕

```
0 0 171.32 10.98 re
W
n
q
0 0 0 rg
0 0 0 RG
0 w
q
107.68 1.15 m
107.68 8.55 l
107.68 8.57 107.68 8.59 107.68 8.6 c
107.67 8.62 107.67 8.63 107.66 8.64 c
107.65 8.65 107.64 8.66 107.63 8.67 c
```


*C:\iText\SurenExamples\results\result.pdf - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
result.pdf

```
endstream
endobj
4 0 obj
<</Contents 5 0 R/MediaBox[0 0 210 200]/Parent 2 0 R/Resources<</Font<</F1 6 0 R>>/XObject<</Fm1 7 0 R>>>>/TrimBox[0 0 2
endobj
1 0 obj
<</Pages 2 0 R/Type/Catalog>>
endobj
3 0 obj
```

Normal text file   length : 102.124   lines : 2.495   Ln : 38   Col : 1   Sel : 0 | 0   Unix (LF)   ANSI   INS

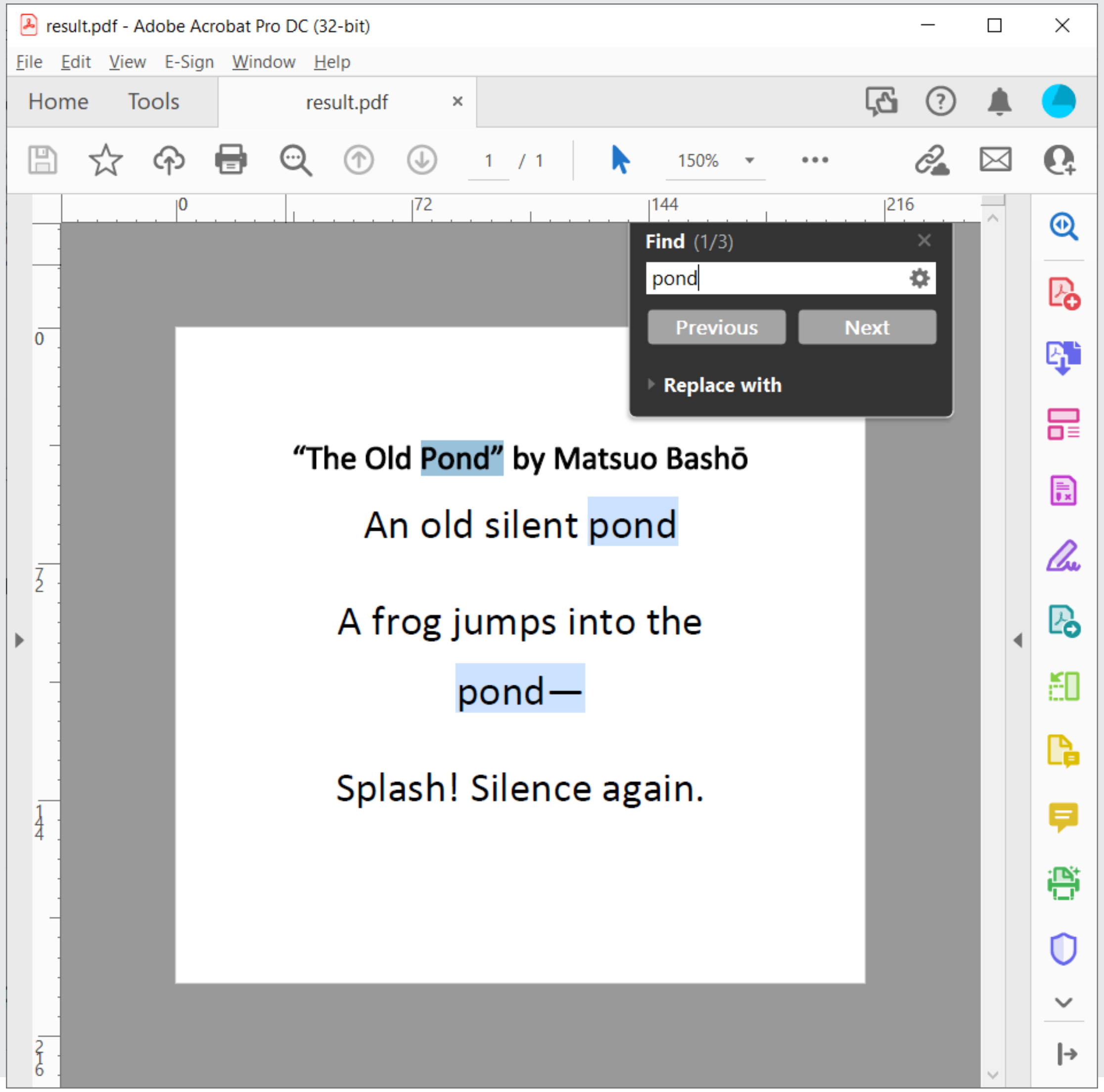
```
7 0 obj
<</BBox[0 0 171.32 10.98]/Length 58/Resources<</XObject<</Fm1 13 0 R>>>>/Subtype/Form/Type/XObject>>stream
1 0 0 -1 0 10.98 cm
% svg root
q
1 0 0 1 0 0 cm
/Fm1 Do
Q
endstream
```

Normal text file   length : 102.124   lines : 2.495   Ln : 298   Col : 7   Sel : 0 | 0   Unix (LF)   ANSI   INS

```
13 0 obj
<</BBox[0 0 171.32 10.98]/Length 65623/Resources<<>>/Subtype/Form/Type/XObject>>stream
0 0 171.32 10.98 re
W
n
q
0 0 0 rg
0 0 0 RG
0 w
q
% path
107.68 1.15 m
107.68 8.55 l
107.68 8.57 107.68 8.59 107.68 8.6 c
107.67 8.62 107.67 8.63 107.66 8.64 c
107.65 8.65 107.64 8.66 107.63 8.67 c
107.62 8.68 107.6 8.69 v
```

Normal text file   length : 102.124   lines : 2.495   Ln : 364   Col : 1   Sel : 0 | 0   Unix (LF)   ANSI   INS

Case 2 - Solution



result.pdf - Adobe Acrobat Pro DC (32-bit)
File Edit View E-Sign Window Help
Home Tools result.pdf ×
1 / 1
150%
Find (1/3)
pond
Previous
Next
Replace with

"The Old Pond" by Matsuo Bashō

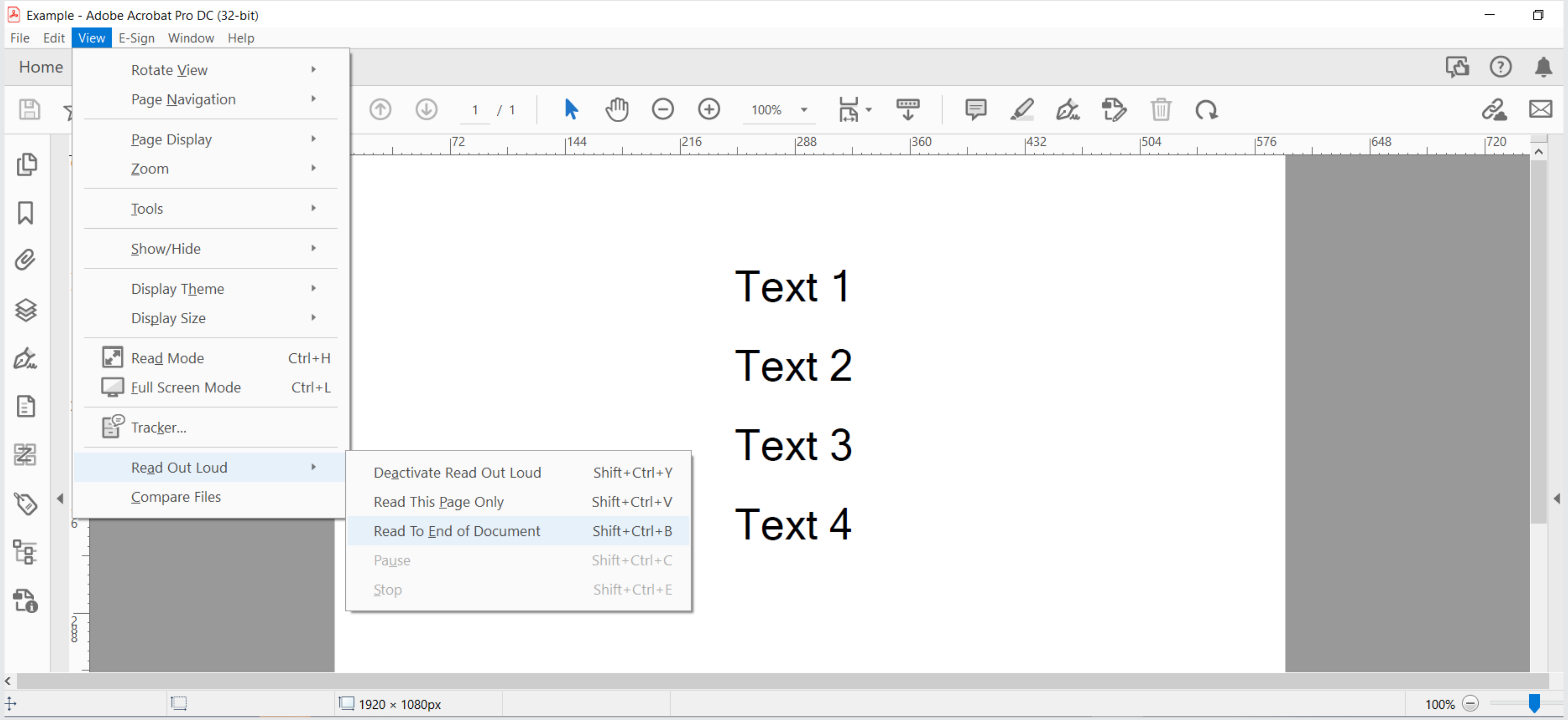
An old silent pond

A frog jumps into the

pond—

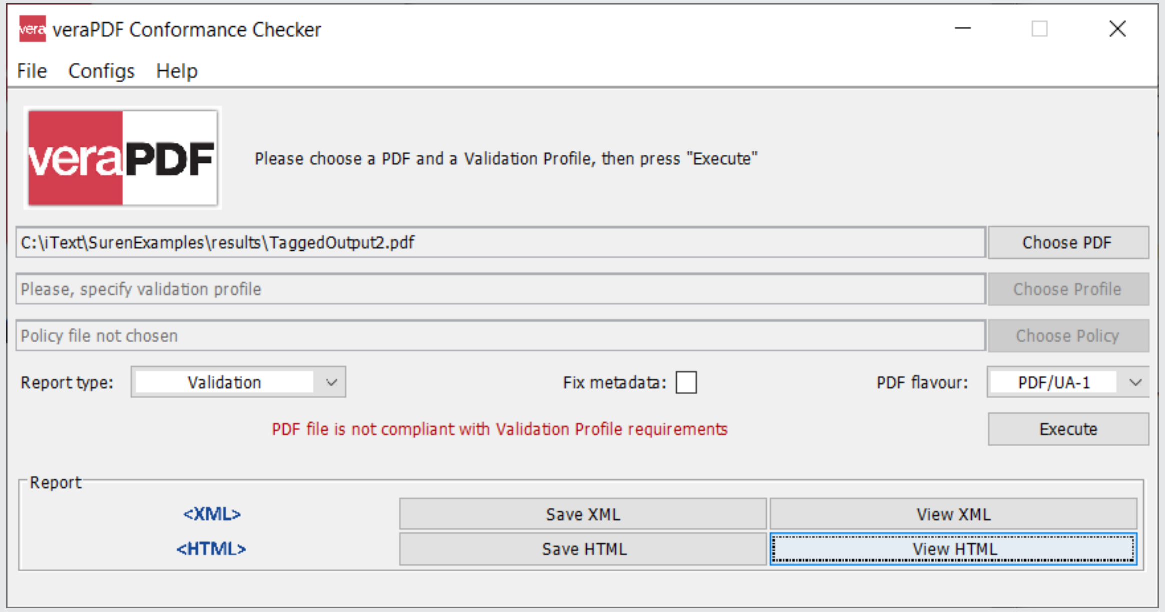
Splash! Silence again.

Case 3 - Usecase



Case 3 - Debugging

working together!



Please choose a PDF and a Validation Profile, then press "Execute"

C:\iText\SurenExamples\results\TaggedOutput2.pdf

Please, specify validation profile

Policy file not chosen

Report type: Validation

Fix metadata: □

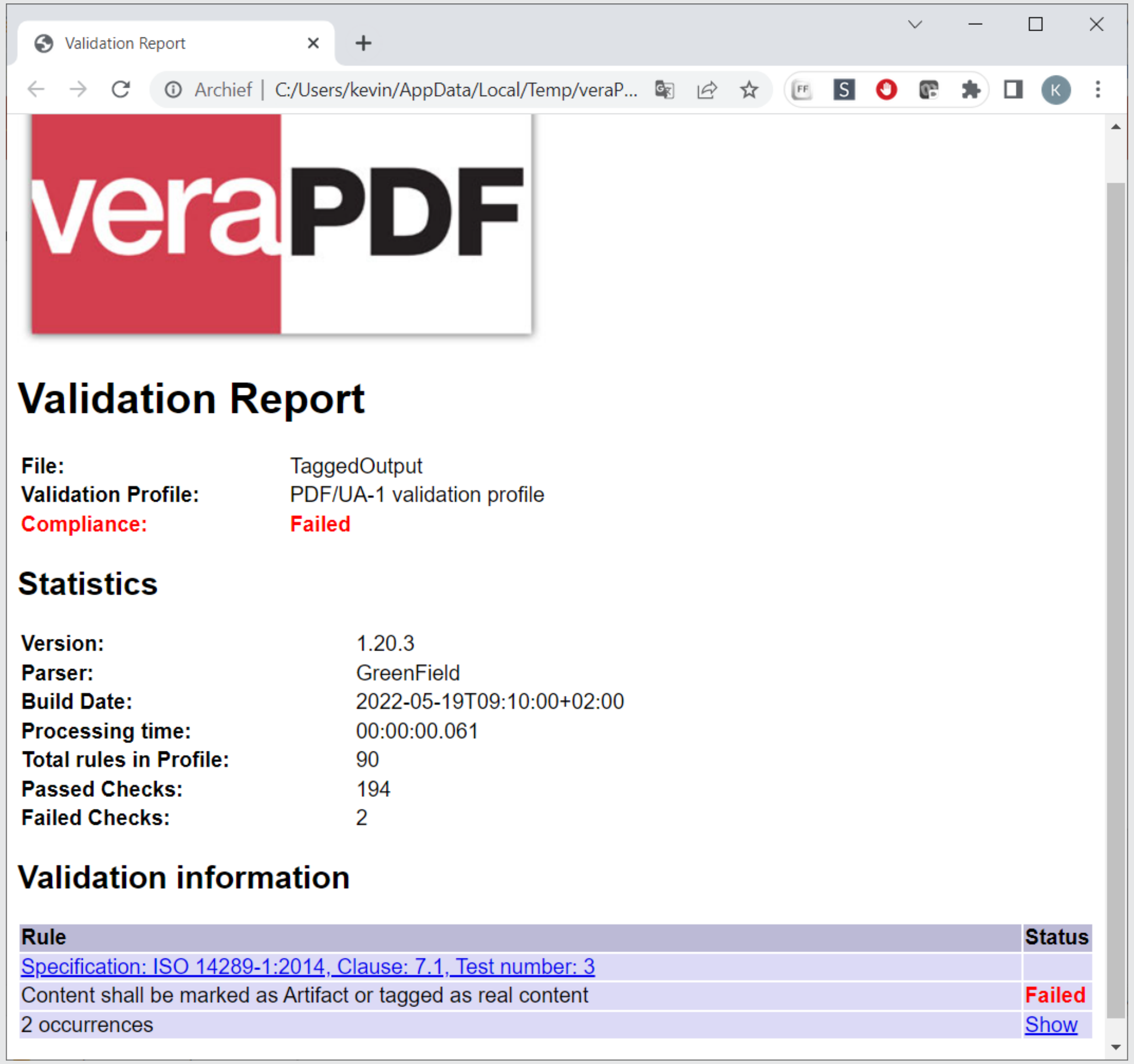
PDF flavour: PDF/UA-1

PDF file is not compliant with Validation Profile requirements

Report

<XML>

<HTML>



Validation Report

File:	TaggedOutput
Validation Profile:	PDF/UA-1 validation profile
Compliance:	Failed

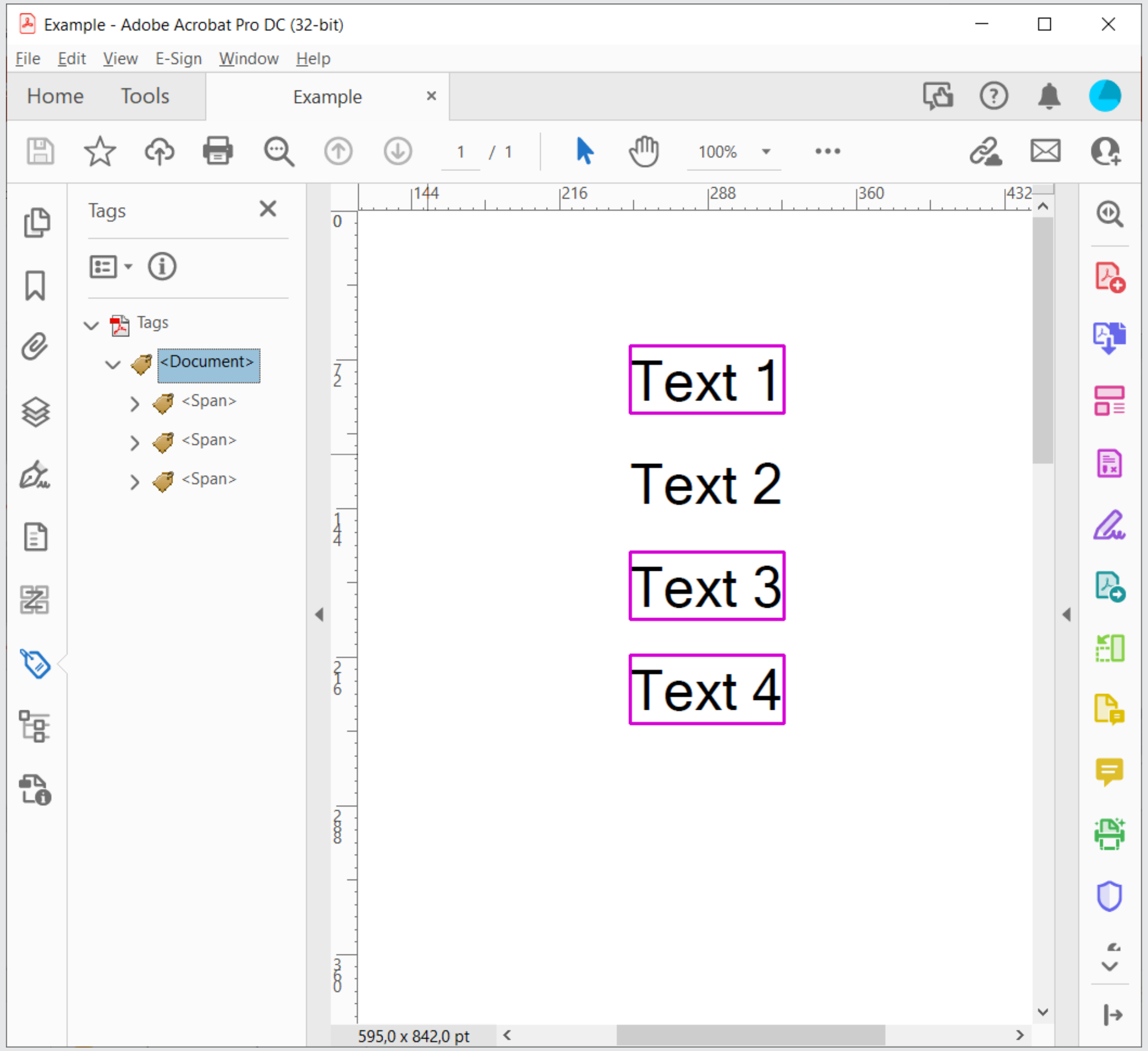
Statistics

Version:	1.20.3
Parser:	GreenField
Build Date:	2022-05-19T09:10:00+02:00
Processing time:	00:00:00.061
Total rules in Profile:	90
Passed Checks:	194
Failed Checks:	2

Validation information

Rule	Status
Specification: ISO 14289-1:2014, Clause: 7.1, Test number: 3	
Content shall be marked as Artifact or tagged as real content	Failed
2 occurrences	Show

Case 3 - Debugging



Example - Adobe Acrobat Pro DC (32-bit)

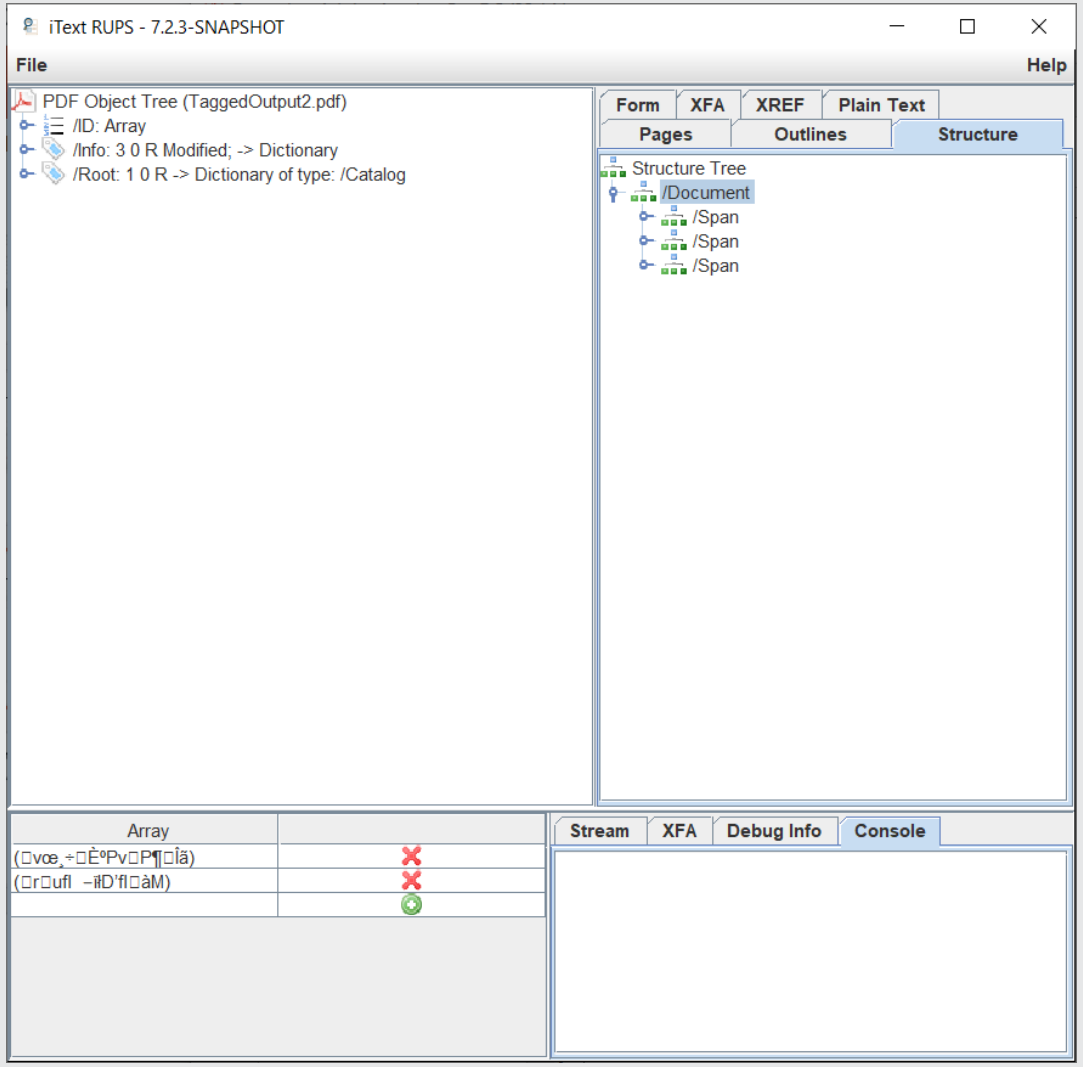
- Tags
  - <Document>
    - <Span>
    - <Span>
    - <Span>

Text 1

Text 2

Text 3

Text 4



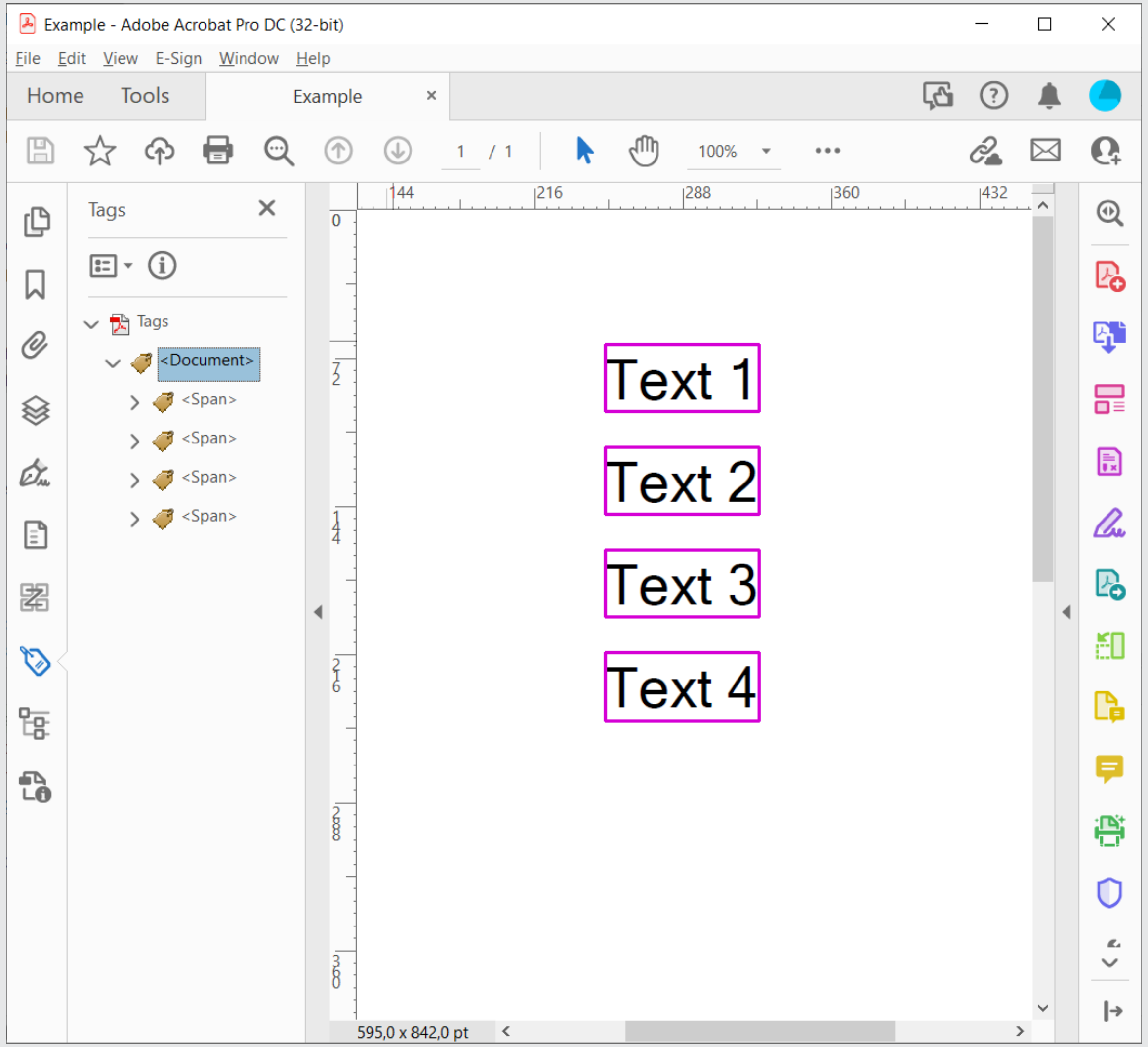
iText RUPS - 7.2.3-SNAPSHOT

- PDF Object Tree (TaggedOutput2.pdf)
  - /ID: Array
  - /Info: 3 0 R Modified; -> Dictionary
  - /Root: 1 0 R -> Dictionary of type: /Catalog

- Structure Tree
  - /Document
    - /Span
    - /Span
    - /Span

Array	
(□vœ¸÷□ÈºPv□P¶□Îã)	✗
(□r□ufl –ïłD'fl□àM)	✗
	⊕

Case 3 – Solution?


Example - Adobe Acrobat Pro DC (32-bit)
File Edit View E-Sign Window Help
Home Tools Example
1 / 1
100%
144 216 288 360 432
Tags
595,0 x 842,0 pt

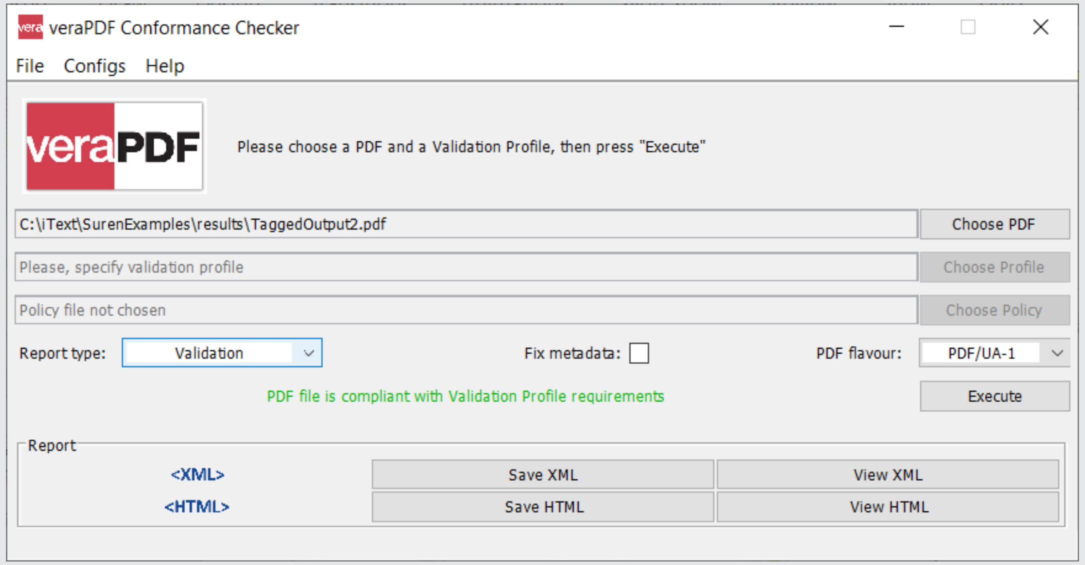
```
Tags
  <Document>
    <Span>
    <Span>
    <Span>
    <Span>
```

Text 1

Text 2

Text 3

Text 4



Please choose a PDF and a Validation Profile, then press "Execute"

C:\iText\SurenExamples\results\TaggedOutput2.pdf

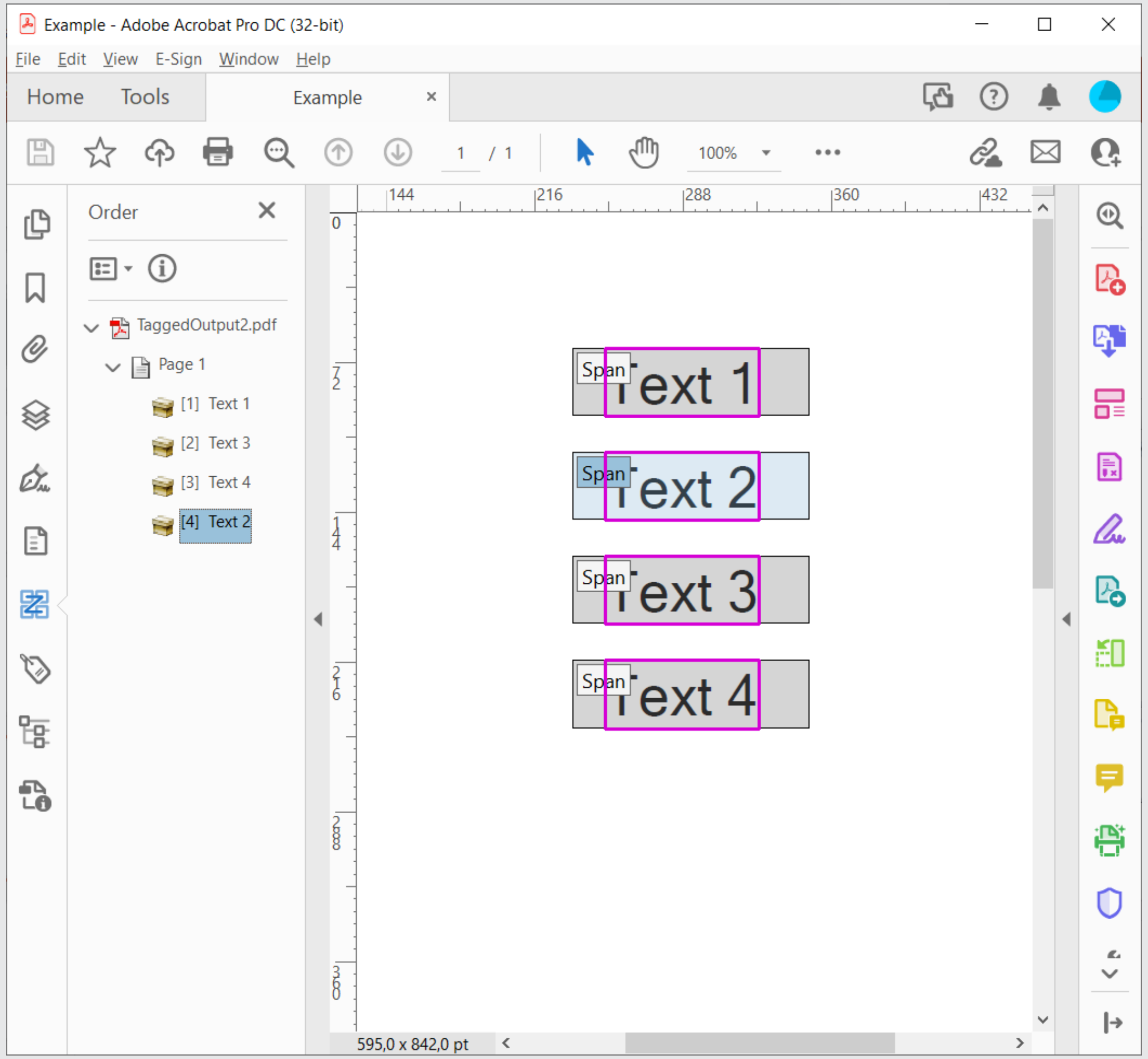
Please, specify validation profile

Policy file not chosen

Report type: Validation   Fix metadata: □   PDF flavour: PDF/UA-1

PDF file is compliant with Validation Profile requirements

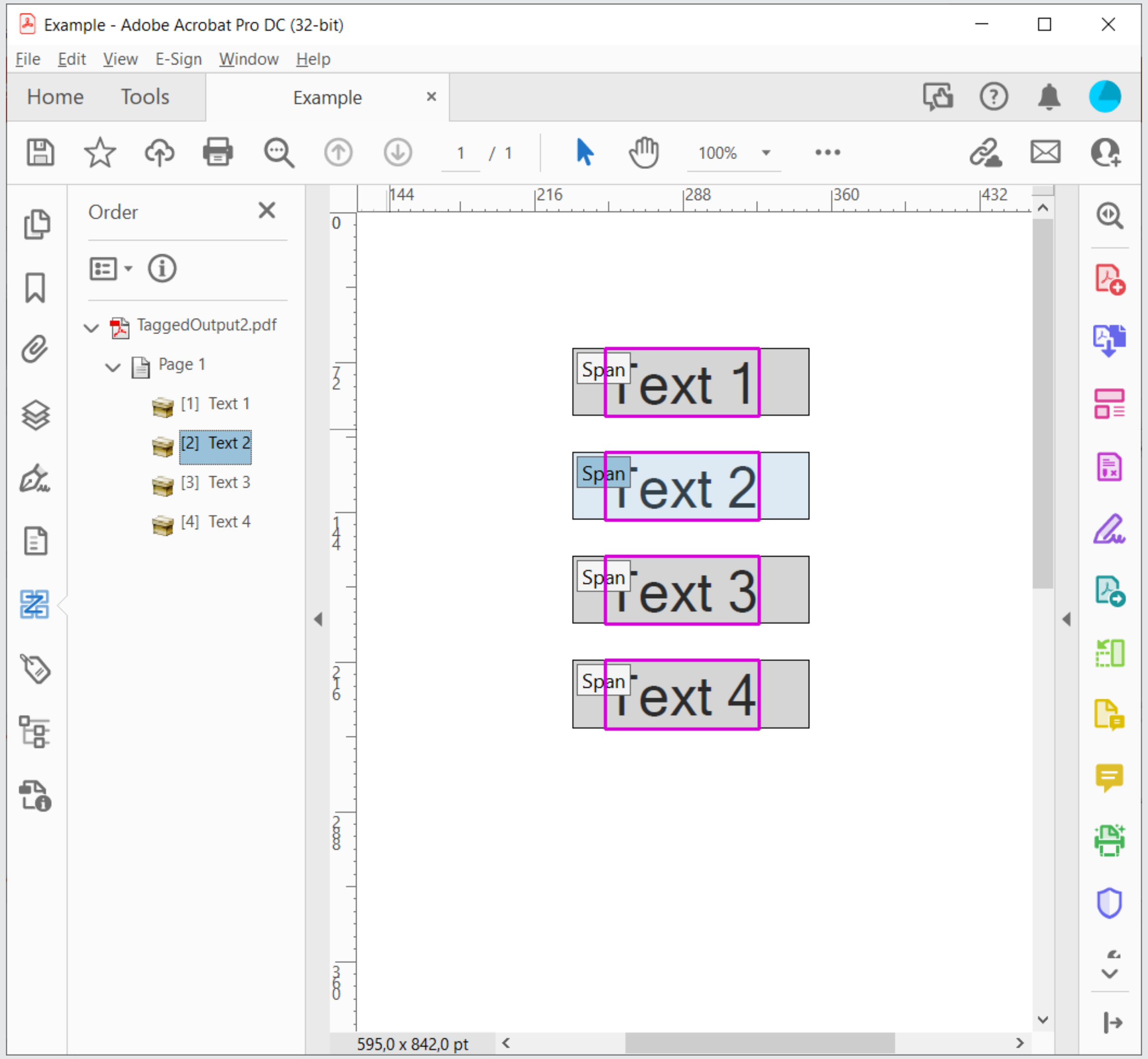
Case 3 – Solution!



Example - Adobe Acrobat Pro DC (32-bit)

Order

- TaggedOutput2.pdf
  - Page 1
    - [1] Text 1
    - [2] Text 3
    - [3] Text 4
    - [4] Text 2



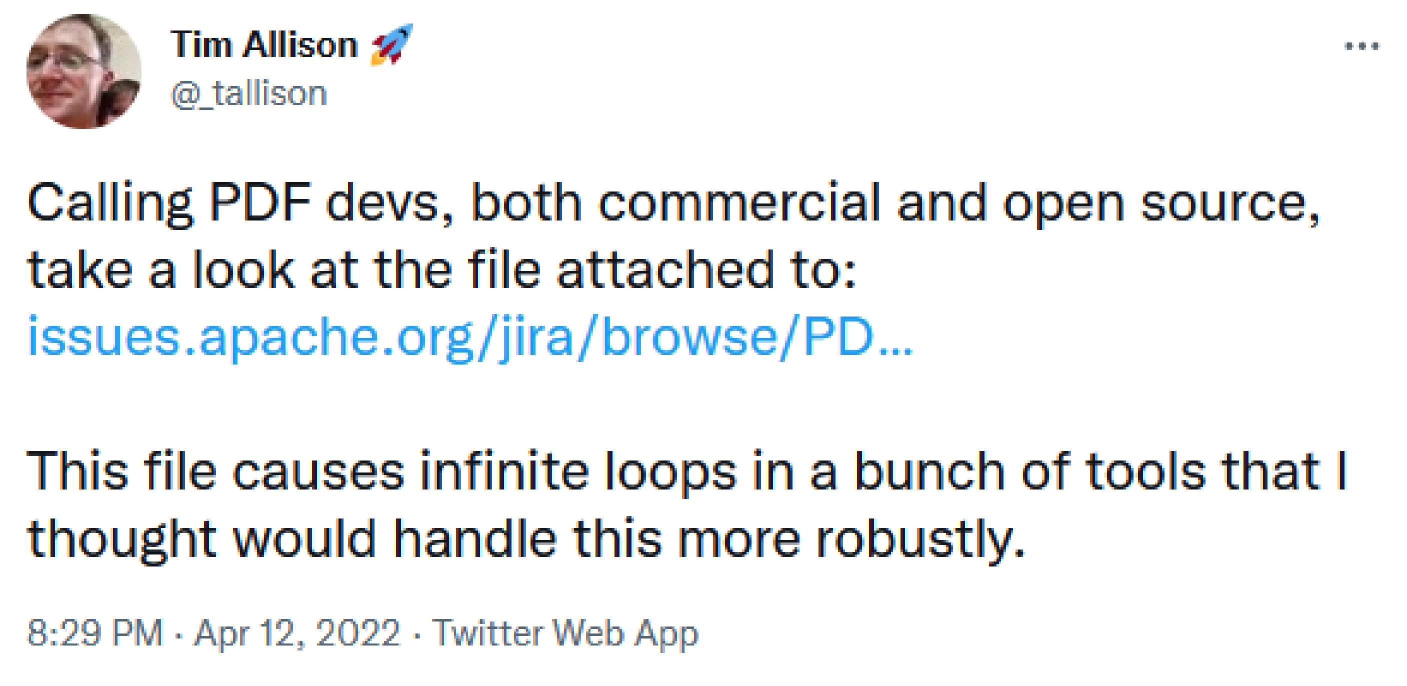
Example - Adobe Acrobat Pro DC (32-bit)

Order

- TaggedOutput2.pdf
  - Page 1
    - [1] Text 1
    - [2] Text 2
    - [3] Text 3
    - [4] Text 4

Case 4

working
together!



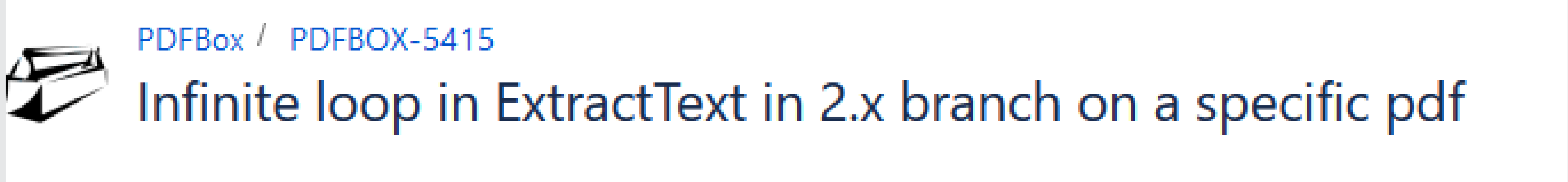
Tim Allison 🚀
@_tallison

Calling PDF devs, both commercial and open source, take a look at the file attached to: issues.apache.org/jira/browse/PD...

This file causes infinite loops in a bunch of tools that I thought would handle this more robustly.

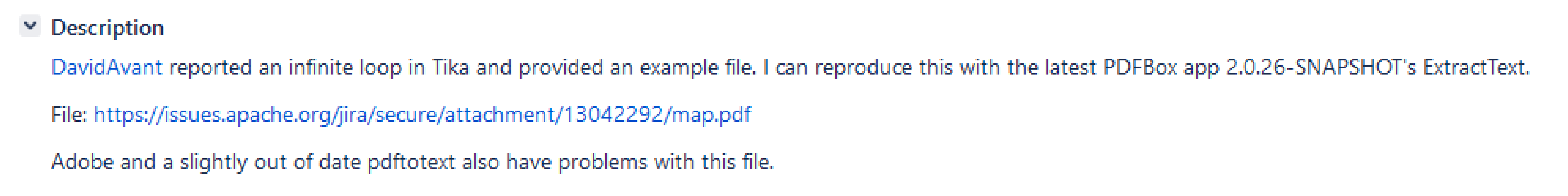
8:29 PM · Apr 12, 2022 · Twitter Web App

Case 4 – Witness Report



PDFBox / PDFBOX-5415

Infinite loop in ExtractText in 2.x branch on a specific pdf



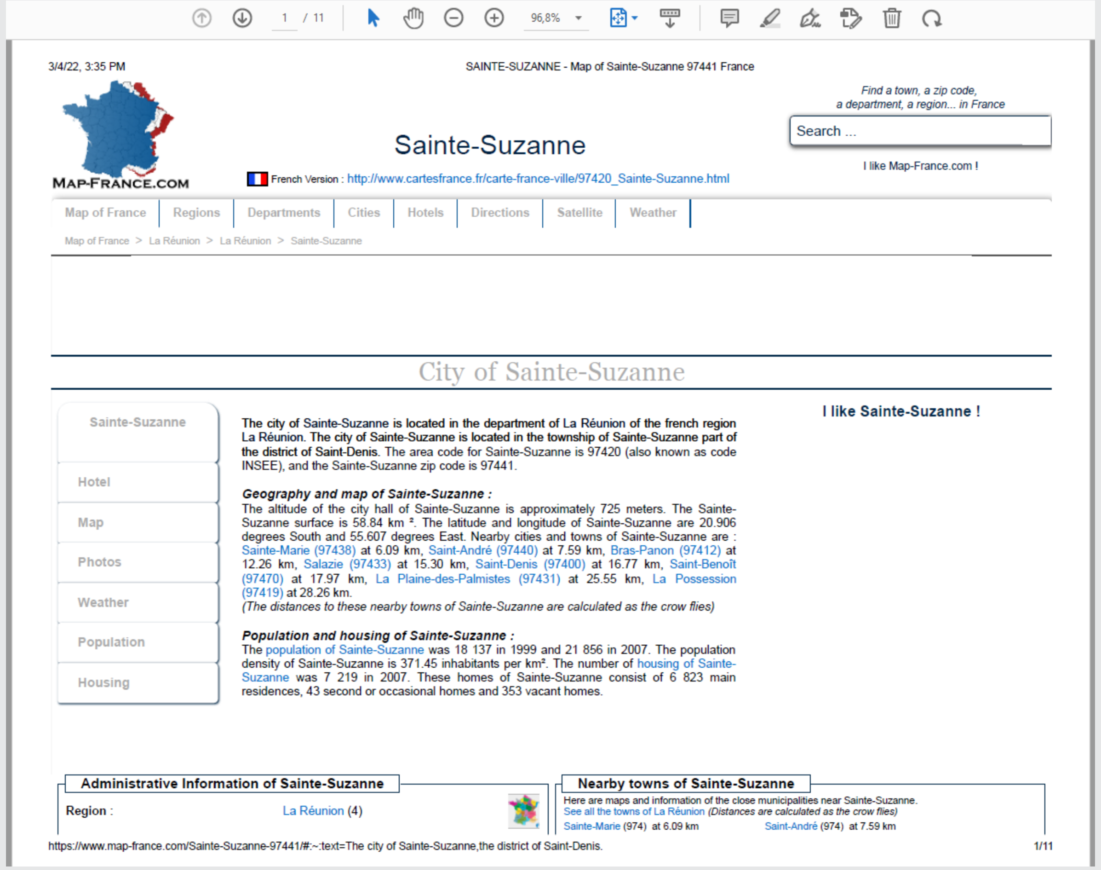
Description

DavidAvant reported an infinite loop in Tika and provided an example file. I can reproduce this with the latest PDFBox app 2.0.26-SNAPSHOT's ExtractText.

File: https://issues.apache.org/jira/secure/attachment/13042292/map.pdf

Adobe and a slightly out of date pdftotext also have problems with this file.

Case 4 – The initial look



3/4/22, 3:35 PM

SAINTE-SUZANNE - Map of Sainte-Suzanne 97441 France

Find a town, a zip code,
a department, a region... in France

I like Map-France.com !

Sainte-Suzanne

French Version : http://www.cartesfrance.fr/carte-france-ville/97420_Sainte-Suzanne.html

Map of France | Regions | Departments | Cities | Hotels | Directions | Satellite | Weather
Map of France > La Réunion > La Réunion > Sainte-Suzanne

City of Sainte-Suzanne

Sainte-Suzanne
Hotel
Map
Photos
Weather
Population
Housing

The city of Sainte-Suzanne is located in the department of La Réunion of the french region La Réunion. The city of Sainte-Suzanne is located in the township of Sainte-Suzanne part of the district of Saint-Denis. The area code for Sainte-Suzanne is 97420 (also known as code INSEE), and the Sainte-Suzanne zip code is 97441.

Geography and map of Sainte-Suzanne :
The altitude of the city hall of Sainte-Suzanne is approximately 725 meters. The Sainte-Suzanne surface is 58.84 km ². The latitude and longitude of Sainte-Suzanne are 20.906 degrees South and 55.607 degrees East. Nearby cities and towns of Sainte-Suzanne are : Sainte-Marie (97438) at 6.09 km, Saint-André (97440) at 7.59 km, Bras-Panon (97412) at 12.26 km, Salazie (97433) at 15.30 km, Saint-Denis (97400) at 16.77 km, Saint-Benoît (97470) at 17.97 km, La Plaine-des-Palmistes (97431) at 25.55 km, La Possession (97419) at 28.26 km.
(The distances to these nearby towns of Sainte-Suzanne are calculated as the crow flies)

Population and housing of Sainte-Suzanne :
The population of Sainte-Suzanne was 18 137 in 1999 and 21 856 in 2007. The population density of Sainte-Suzanne is 371.45 inhabitants per km². The number of housing of Sainte-Suzanne was 7 219 in 2007. These homes of Sainte-Suzanne consist of 6 823 main residences, 43 second or occasional homes and 353 vacant homes.

I like Sainte-Suzanne !

Administrative Information of Sainte-Suzanne

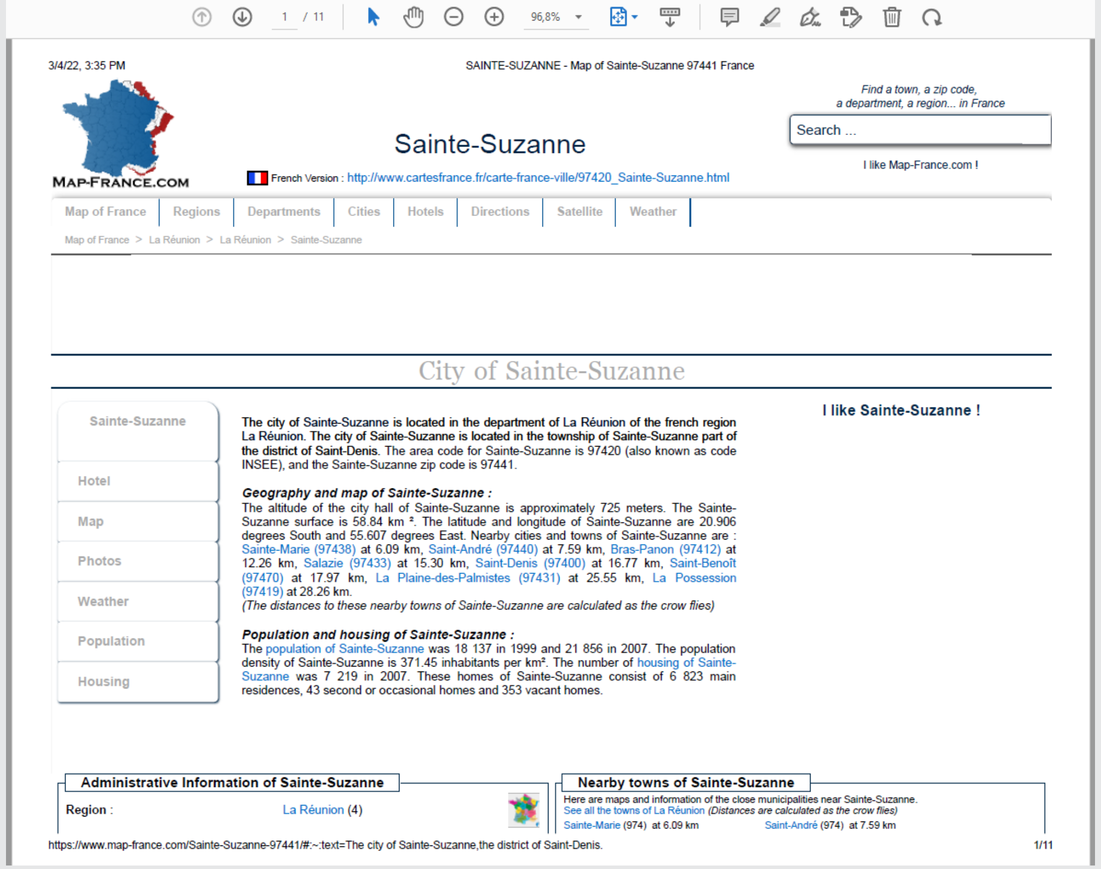
Region : La Réunion (4)

Nearby towns of Sainte-Suzanne

Here are maps and information of the close municipalities near Sainte-Suzanne.
See all the towns of La Réunion *(Distances are calculated as the crow flies)*
Sainte-Marie (974) at 6.09 km Saint-André (974) at 7.59 km

https://www.map-france.com/Sainte-Suzanne-97441/#:~:text=The city of Sainte-Suzanne,the district of Saint-Denis. 1/11

Case 4 – The initial look



3/4/22, 3:35 PM

SAINTE-SUZANNE - Map of Sainte-Suzanne 97441 France

Find a town, a zip code,
a department, a region... in France

Sainte-Suzanne

French Version : http://www.cartesfrance.fr/carte-france-ville/97420_Sainte-Suzanne.html

I like Map-France.com !

City of Sainte-Suzanne

I like Sainte-Suzanne !

The city of Sainte-Suzanne is located in the department of La Réunion of the french region La Réunion. The city of Sainte-Suzanne is located in the township of Sainte-Suzanne part of the district of Saint-Denis. The area code for Sainte-Suzanne is 97420 (also known as code INSEE), and the Sainte-Suzanne zip code is 97441.

Geography and map of Sainte-Suzanne :
The altitude of the city hall of Sainte-Suzanne is approximately 725 meters. The Sainte-Suzanne surface is 58.84 km ². The latitude and longitude of Sainte-Suzanne are 20.906 degrees South and 55.607 degrees East. Nearby cities and towns of Sainte-Suzanne are : Sainte-Marie (97438) at 6.09 km, Saint-André (97440) at 7.59 km, Bras-Panon (97412) at 12.26 km, Salazie (97433) at 15.30 km, Saint-Denis (97400) at 16.77 km, Saint-Benoît (97470) at 17.97 km, La Plaine-des-Palmistes (97431) at 25.55 km, La Possession (97419) at 28.26 km.
(The distances to these nearby towns of Sainte-Suzanne are calculated as the crow flies)

Population and housing of Sainte-Suzanne :
The population of Sainte-Suzanne was 18 137 in 1999 and 21 856 in 2007. The population density of Sainte-Suzanne is 371.45 inhabitants per km². The number of housing of Sainte-Suzanne was 7 219 in 2007. These homes of Sainte-Suzanne consist of 6 823 main residences, 43 second or occasional homes and 353 vacant homes.

Administrative Information of Sainte-Suzanne

Region : La Réunion (4)

Nearby towns of Sainte-Suzanne

Here are maps and information of the close municipalities near Sainte-Suzanne.
See all the towns of La Réunion *(Distances are calculated as the crow flies)*
Sainte-Marie (974) at 6.09 km Saint-André (974) at 7.59 km

https://www.map-france.com/Sainte-Suzanne-97441/#:~:text=The city of Sainte-Suzanne,the district of Saint-Denis. 1/11

It crashes the viewer

Case 4- Break out the tools

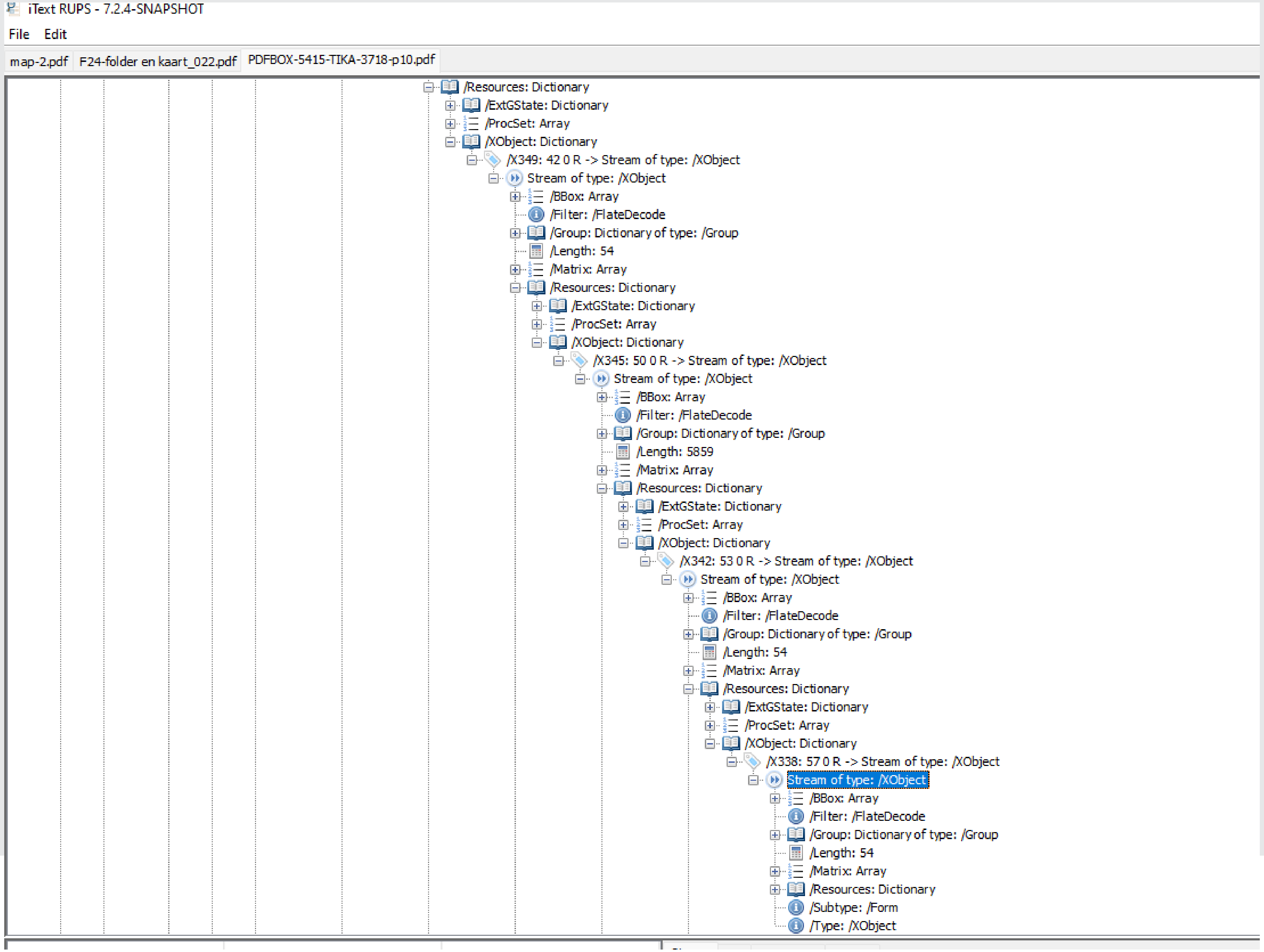
working together!

- Reproduce the issue as-is
  - Run PDF Box code
  - Encounter reported behavior

Case 4 – Switch up the tools

- Run the same code on iText 7
  - Is the bug only in PDF Box or is it a file issue
  - Encounter the same behavior

Case 4 – Dissecting the body



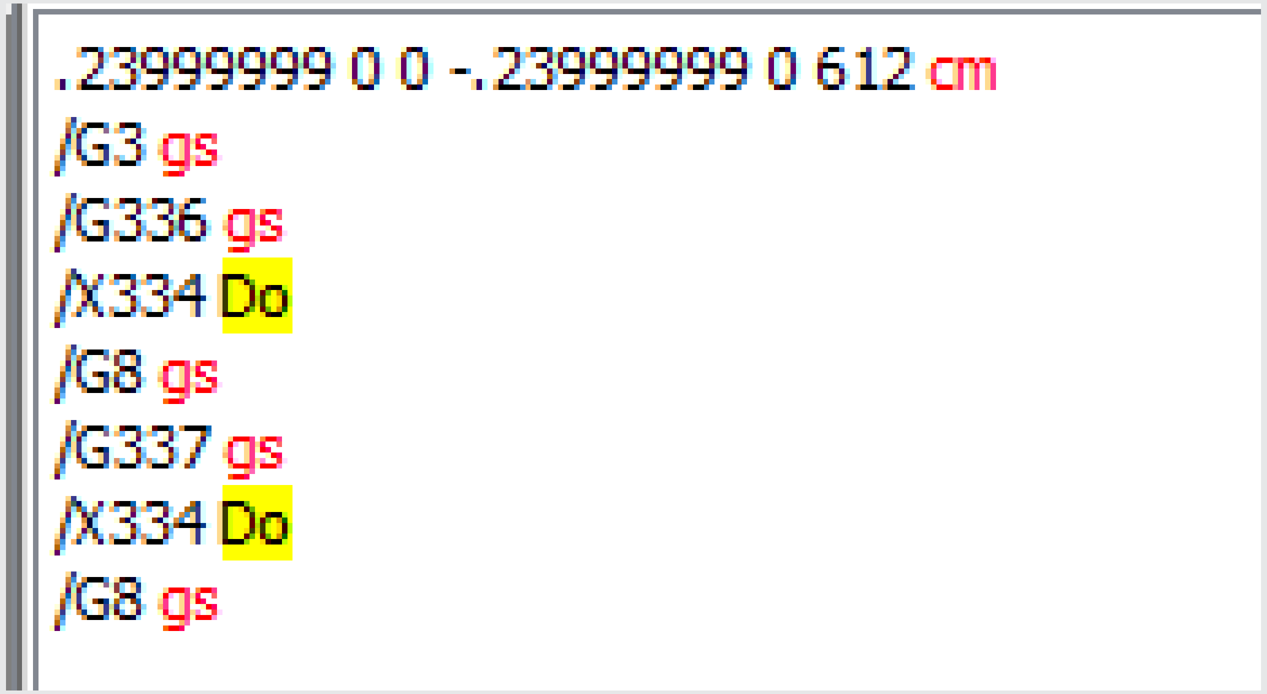
iText RUPS - 7.2.4-SNAPSHOT

File Edit

map-2.pdf F24-folder en kaart_022.pdf PDFBOX-5415-TIKA-3718-p10.pdf

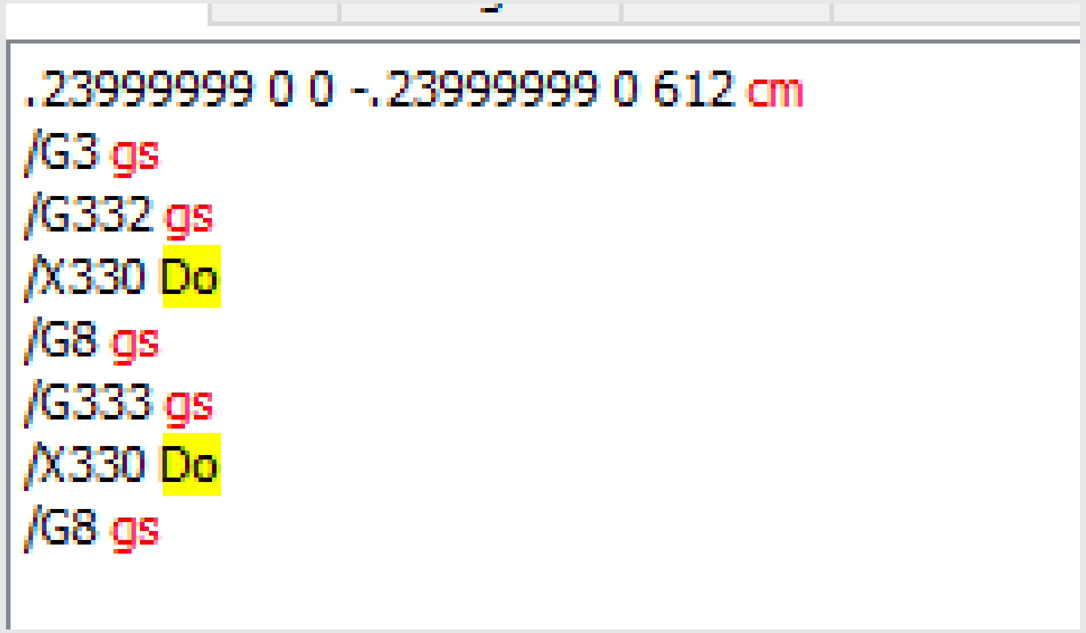
```
/Resources: Dictionary
  /ExtGState: Dictionary
  /ProcSet: Array
  /XObject: Dictionary
    /X349: 42 0 R -> Stream of type: /XObject
      Stream of type: /XObject
        /BBox: Array
        /Filter: /FlateDecode
        /Group: Dictionary of type: /Group
        /Length: 54
        /Matrix: Array
        /Resources: Dictionary
          /ExtGState: Dictionary
          /ProcSet: Array
          /XObject: Dictionary
            /X345: 50 0 R -> Stream of type: /XObject
              Stream of type: /XObject
                /BBox: Array
                /Filter: /FlateDecode
                /Group: Dictionary of type: /Group
                /Length: 5859
                /Matrix: Array
                /Resources: Dictionary
                  /ExtGState: Dictionary
                  /ProcSet: Array
                  /XObject: Dictionary
                    /X342: 53 0 R -> Stream of type: /XObject
                      Stream of type: /XObject
                        /BBox: Array
                        /Filter: /FlateDecode
                        /Group: Dictionary of type: /Group
                        /Length: 54
                        /Matrix: Array
                        /Resources: Dictionary
                          /ExtGState: Dictionary
                          /ProcSet: Array
                          /XObject: Dictionary
                            /X338: 57 0 R -> Stream of type: /XObject
                              Stream of type: /XObject
                                /BBox: Array
                                /Filter: /FlateDecode
                                /Group: Dictionary of type: /Group
                                /Length: 54
                                /Matrix: Array
                                /Resources: Dictionary
                                /Subtype: /Form
                                /Type: /XObject
```

Case 4 – Examining the Blood Work



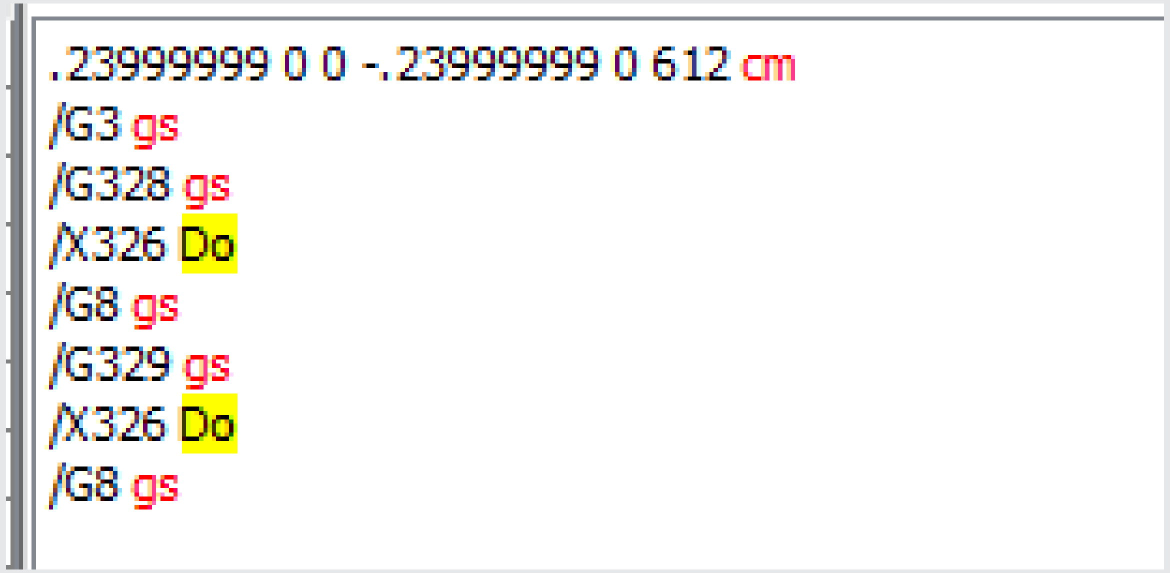
```
.23999999 0 0 -.23999999 0 612 cm
/G3 gs
/G336 gs
/X334 Do
/G8 gs
/G337 gs
/X334 Do
/G8 gs
```

Case 4 – Drilling down into blood results



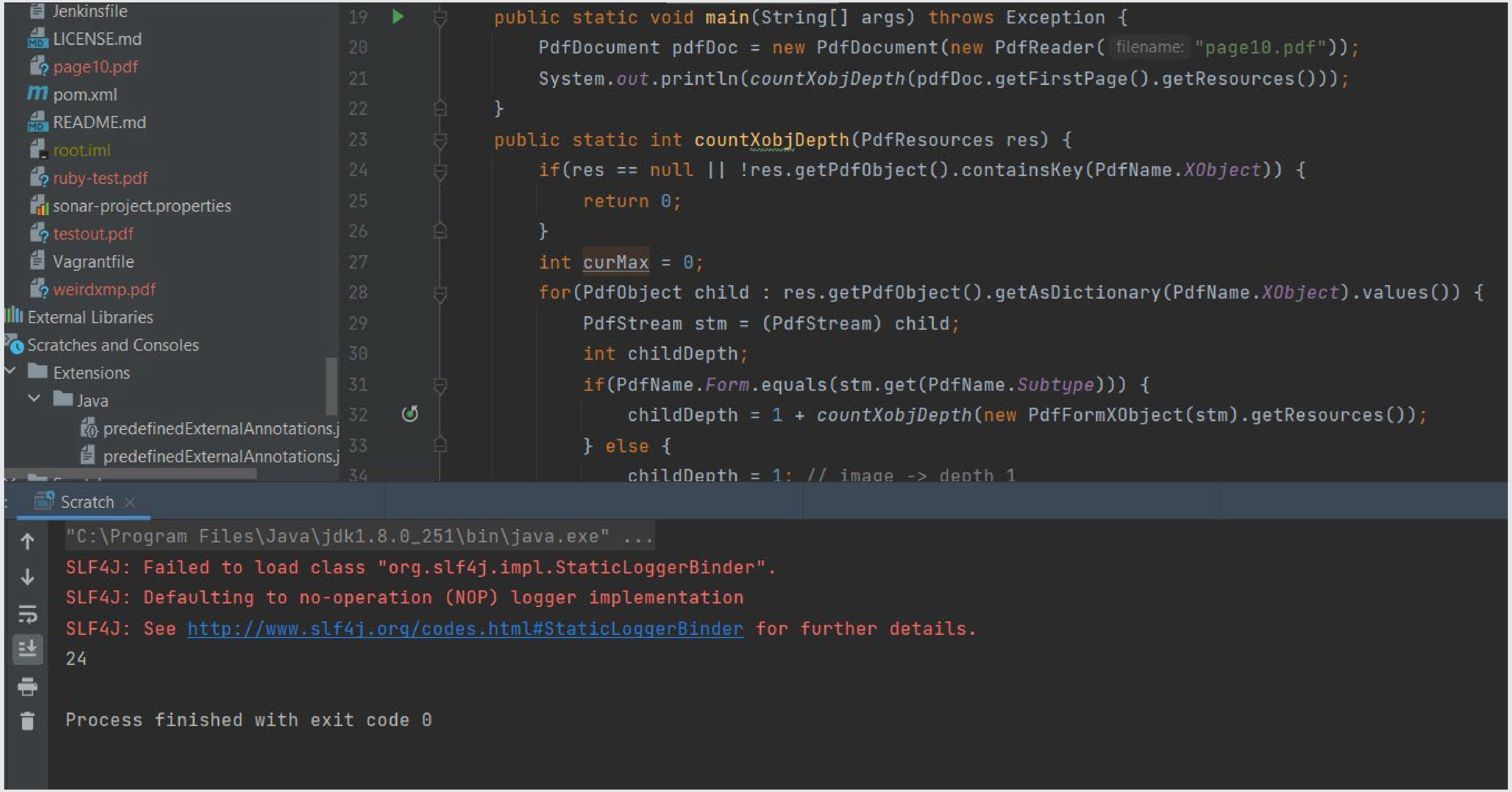
```
.23999999 0 0 -.23999999 0 612 cm
/G3 gs
/G332 gs
/X330 Do
/G8 gs
/G333 gs
/X330 Do
/G8 gs
```

Case 4 – Noticing the oddity



```
.23999999 0 0 -.23999999 0 612 cm
/G3 gs
/G328 gs
/X326 Do
/G8 gs
/G329 gs
/X326 Do
/G8 gs
```

Case 4 – Crunching the numbers



```
public static void main(String[] args) throws Exception {
    PdfDocument pdfDoc = new PdfDocument(new PdfReader( filename: "page10.pdf"));
    System.out.println(countXobjDepth(pdfDoc.getFirstPage().getResources()));
}
public static int countXobjDepth(PdfResources res) {
    if(res == null || !res.getPdfObject().containsKey(PdfName.XObject)) {
        return 0;
    }
    int curMax = 0;
    for(PdfObject child : res.getPdfObject().getAsDictionary(PdfName.XObject).values()) {
        PdfStream stm = (PdfStream) child;
        int childDepth;
        if(PdfName.Form.equals(stm.get(PdfName.Subtype))) {
            childDepth = 1 + countXobjDepth(new PdfFormXObject(stm).getResources());
        } else {
            childDepth = 1; // image -> depth 1
```

```
"C:\Program Files\Java\jdk1.8.0_251\bin\java.exe" ...
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
24

Process finished with exit code 0
```

Case 4 – Conclusion

- The PDF file isn’t broken
- Implementations aren’t broken

- PDF files can have complex and intricate constructs
  - Flexibility of the spec

- Parsers need to find ways around this
  - Cache XObjects
  - Lazy Loading of XObjects

Case 5 – The One with Crowd-sourcing

- Customer reports “broken text extraction”
    - The text was perfectly legible in viewers

Case 5

- Customer reports “broken text extraction”
  - The text was perfectly legible in viewers
  - When extracting, only some letters would be incorrect
    - Yet, consistent incorrect

Case 5

- Customer reports “broken text extraction”
  - The text was perfectly legible in viewers
  - When extracting, only some letters would be incorrect
    - Yet, consistent incorrect
  - Across tools
    - So, not an iText issue

Case 5

- Customer reports “broken text extraction”
  - The text was perfectly legible in viewers
  - When extracting, only some letters would be incorrect
    - Yet, consistent incorrect
  - Across tools
    - So, not an iText issue
  - Through the same producer

Case 5 – Resolution

- The mapping was completely off

- Erroneous behavior of that producer?
- Obfuscation?

working
together!

Mission Report

PDF
association