working together!

PDF Days Europe 2022 | Berlin

PDF Optimization *Horror* Stories

Grizzly PDF autopsies

Datalogics

working together!

- **The Adobe PDF Library** – built on Adobe source code
- **Live support** from PDF technology experts
- **SDKs and command-line tools** for building large applications – great for OEMs, system integrators and enterprise developers

www.datalogics.com

What makes Processing PDFs into *Horror* stories?

- Ballooning file size
- Hangs /infinite-loops

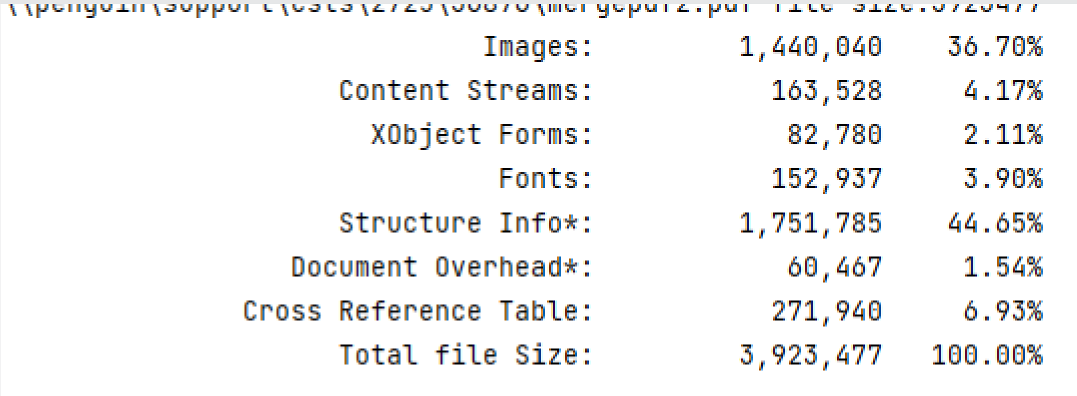
- Most issues are edge-cases rather than Horror Stories

Sub-optimal StructureTrees

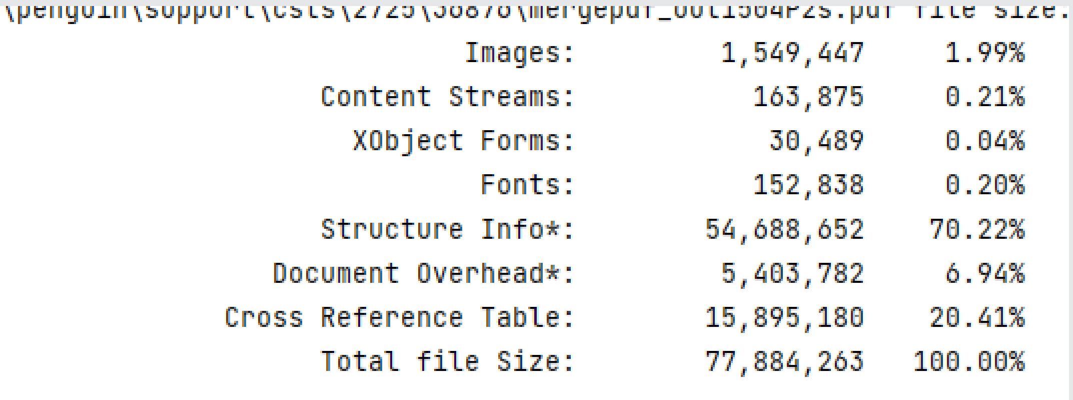
working together!

- Case 1: The structure tree that grew when it was copied.
- Case 2: Phantom Limbs in the StructureTree

36876: the StructureTree that grew…

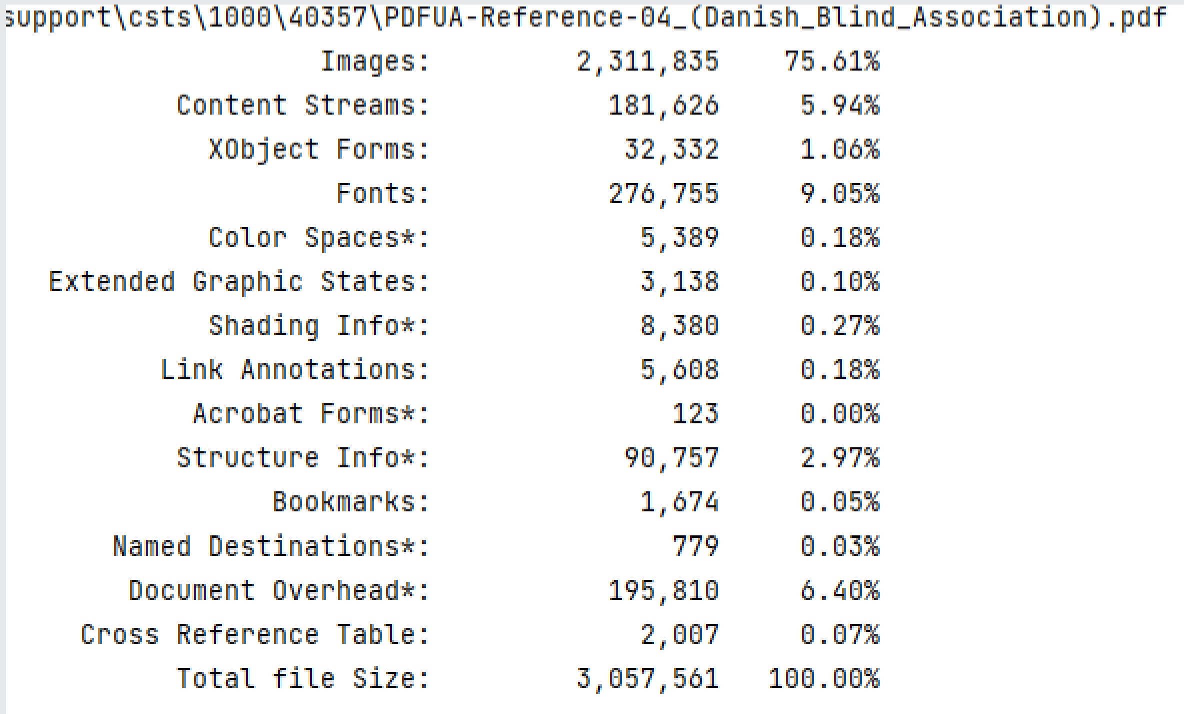


```
\\penguin\support\csts\2725\36876\mergepdf2.pdf file size:3923477
                    Images:        1,440,040    36.70%
           Content Streams:          163,528     4.17%
             XObject Forms:           82,780     2.11%
                     Fonts:          152,937     3.90%
           Structure Info*:        1,751,785    44.65%
        Document Overhead*:           60,467     1.54%
     Cross Reference Table:          271,940     6.93%
           Total file Size:        3,923,477   100.00%
```

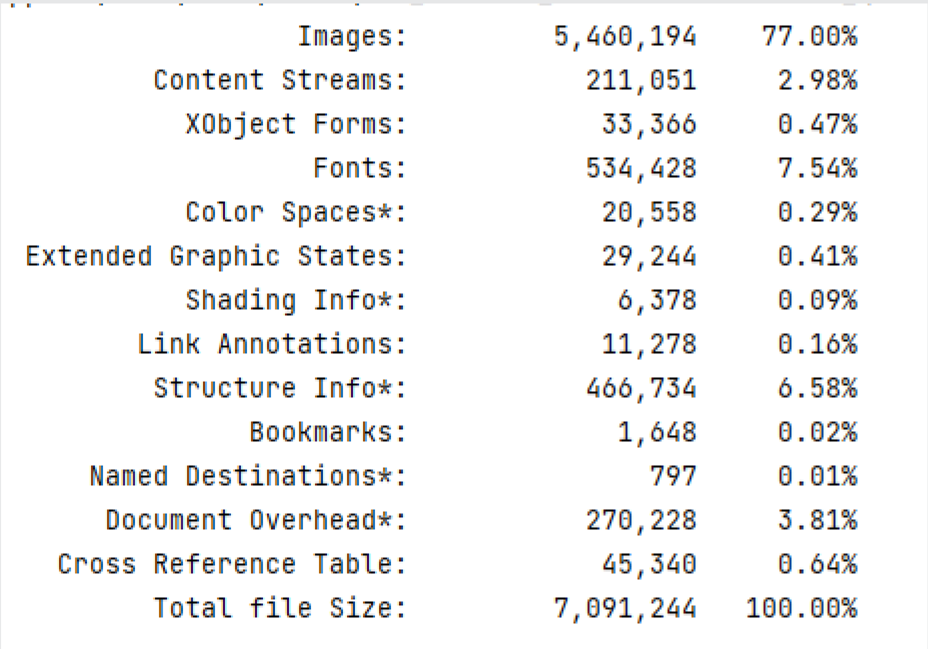


```
\penguin\support\csts\2725\36876\mergepdf_out1504P2s.pdf file size:
                    Images:        1,549,447     1.99%
           Content Streams:          163,875     0.21%
             XObject Forms:           30,489     0.04%
                     Fonts:          152,838     0.20%
           Structure Info*:       54,688,652    70.22%
        Document Overhead*:        5,403,782     6.94%
     Cross Reference Table:       15,895,180    20.41%
           Total file Size:       77,884,263   100.00%
```

40357: Phantom Limbs in StructureTree



```
support\csts\1000\40357\PDFUA-Reference-04_(Danish_Blind_Association).pdf
                     Images:         2,311,835     75.61%
            Content Streams:           181,626      5.94%
              XObject Forms:            32,332      1.06%
                      Fonts:           276,755      9.05%
              Color Spaces*:             5,389      0.18%
   Extended Graphic States:              3,138      0.10%
              Shading Info*:             8,380      0.27%
           Link Annotations:             5,608      0.18%
              Acrobat Forms*:              123      0.00%
            Structure Info*:            90,757      2.97%
                  Bookmarks:             1,674      0.05%
        Named Destinations*:               779      0.03%
         Document Overhead*:           195,810      6.40%
        Cross Reference Table:           2,007      0.07%
            Total file Size:         3,057,561    100.00%
```

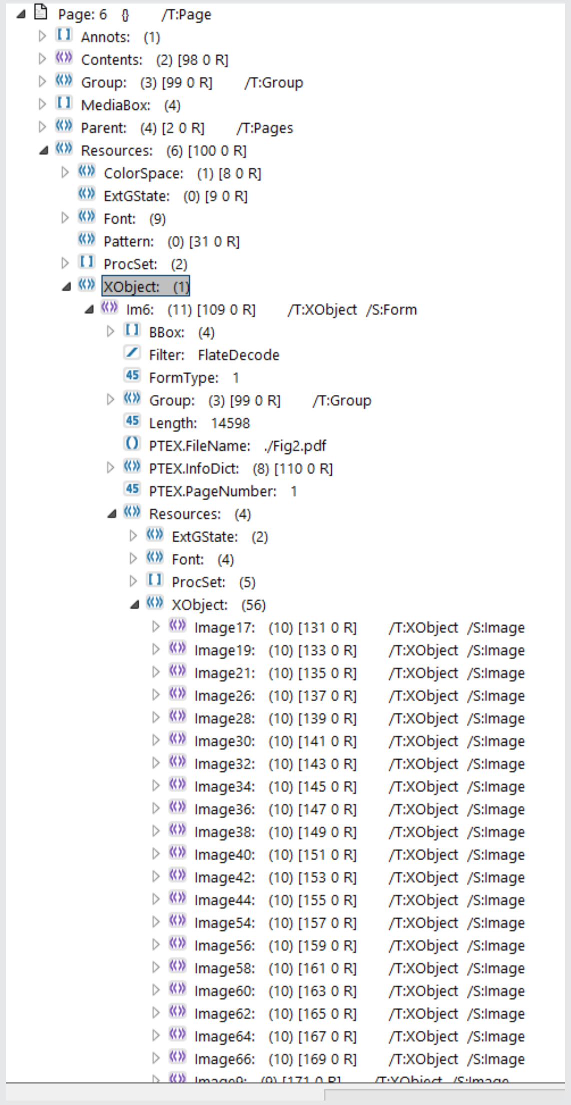


```
                     Images:         5,460,194     77.00%
            Content Streams:           211,051      2.98%
              XObject Forms:            33,366      0.47%
                      Fonts:           534,428      7.54%
              Color Spaces*:            20,558      0.29%
   Extended Graphic States:             29,244      0.41%
              Shading Info*:             6,378      0.09%
           Link Annotations:            11,278      0.16%
            Structure Info*:           466,734      6.58%
                  Bookmarks:             1,648      0.02%
        Named Destinations*:               797      0.01%
         Document Overhead*:           270,228      3.81%
        Cross Reference Table:          45,340      0.64%
            Total file Size:         7,091,244    100.00%
```

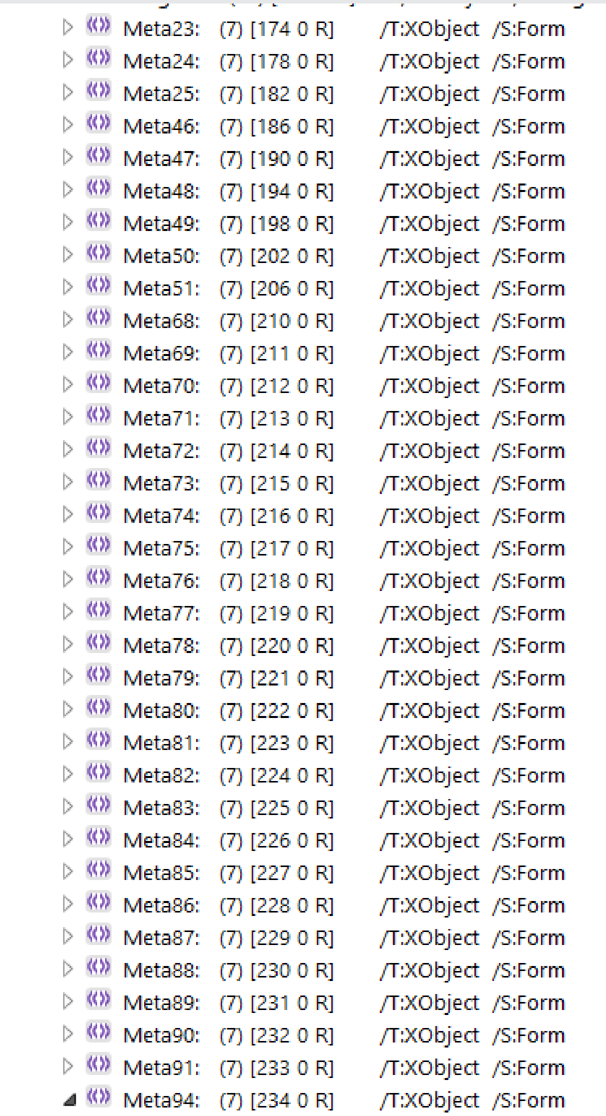
44874: Cyclic resource trees.

- Linearization/Fast Web Preview
  - How it works
  - How it went wrong (for this file)

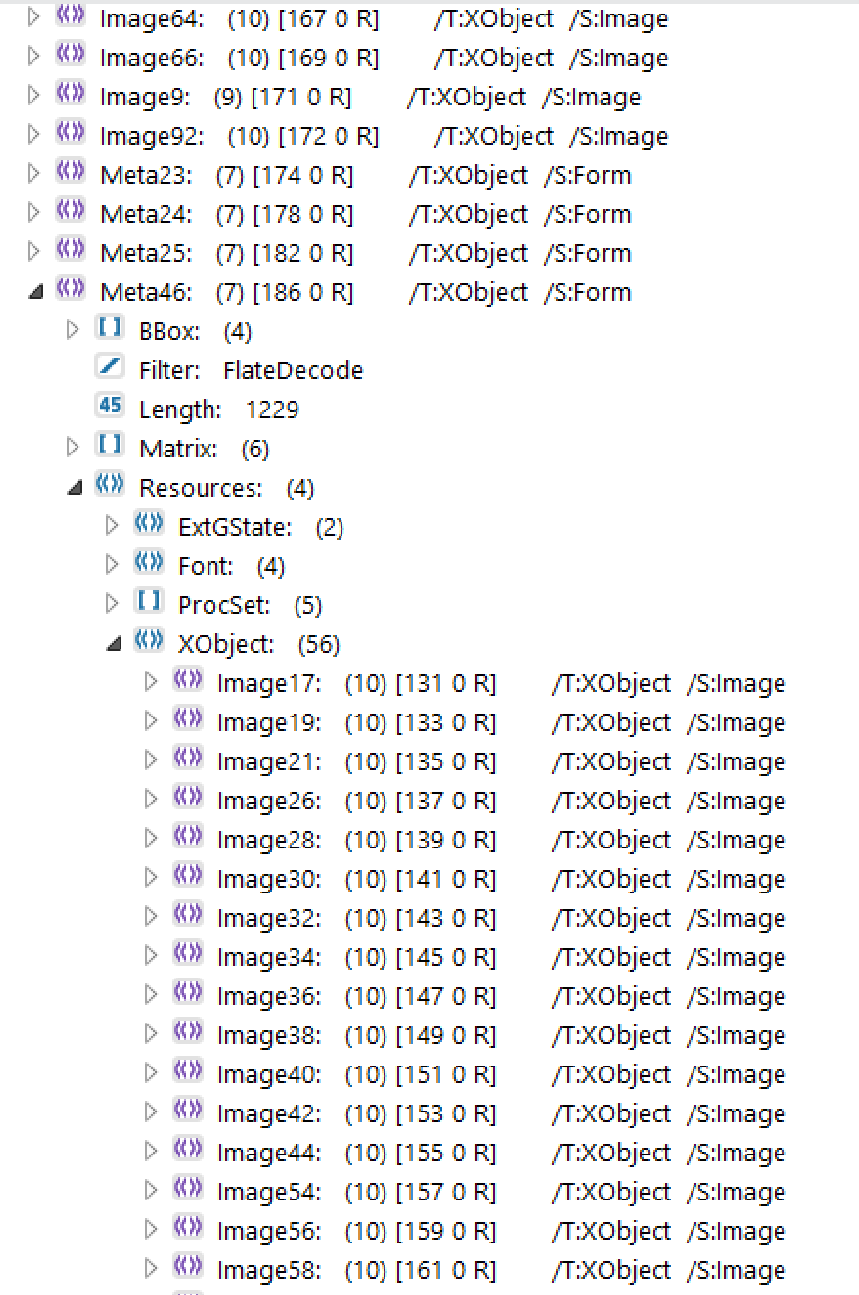
Examining the Resource Tree.



```
Page: 6  {}      /T:Page
  Annots:  (1)
  Contents:  (2) [98 0 R]
  Group:  (3) [99 0 R]      /T:Group
  MediaBox:  (4)
  Parent:  (4) [2 0 R]      /T:Pages
  Resources:  (6) [100 0 R]
    ColorSpace:  (1) [8 0 R]
    ExtGState:  (0) [9 0 R]
    Font:  (9)
    Pattern:  (0) [31 0 R]
    ProcSet:  (2)
    XObject:  (1)
      Im6:  (11) [109 0 R]      /T:XObject  /S:Form
        BBox:  (4)
        Filter:  FlateDecode
        FormType:  1
        Group:  (3) [99 0 R]      /T:Group
        Length:  14598
        PTEX.FileName:  ./Fig2.pdf
        PTEX.InfoDict:  (8) [110 0 R]
        PTEX.PageNumber:  1
        Resources:  (4)
          ExtGState:  (2)
          Font:  (4)
          ProcSet:  (5)
          XObject:  (56)
            Image17:  (10) [131 0 R]      /T:XObject  /S:Image
            Image19:  (10) [133 0 R]      /T:XObject  /S:Image
            Image21:  (10) [135 0 R]      /T:XObject  /S:Image
            Image26:  (10) [137 0 R]      /T:XObject  /S:Image
            Image28:  (10) [139 0 R]      /T:XObject  /S:Image
            Image30:  (10) [141 0 R]      /T:XObject  /S:Image
            Image32:  (10) [143 0 R]      /T:XObject  /S:Image
            Image34:  (10) [145 0 R]      /T:XObject  /S:Image
            Image36:  (10) [147 0 R]      /T:XObject  /S:Image
            Image38:  (10) [149 0 R]      /T:XObject  /S:Image
            Image40:  (10) [151 0 R]      /T:XObject  /S:Image
            Image42:  (10) [153 0 R]      /T:XObject  /S:Image
            Image44:  (10) [155 0 R]      /T:XObject  /S:Image
            Image54:  (10) [157 0 R]      /T:XObject  /S:Image
            Image56:  (10) [159 0 R]      /T:XObject  /S:Image
            Image58:  (10) [161 0 R]      /T:XObject  /S:Image
            Image60:  (10) [163 0 R]      /T:XObject  /S:Image
            Image62:  (10) [165 0 R]      /T:XObject  /S:Image
            Image64:  (10) [167 0 R]      /T:XObject  /S:Image
            Image66:  (10) [169 0 R]      /T:XObject  /S:Image
            Image9:  (9) [171 0 R]      /T:XObject  /S:Image
```



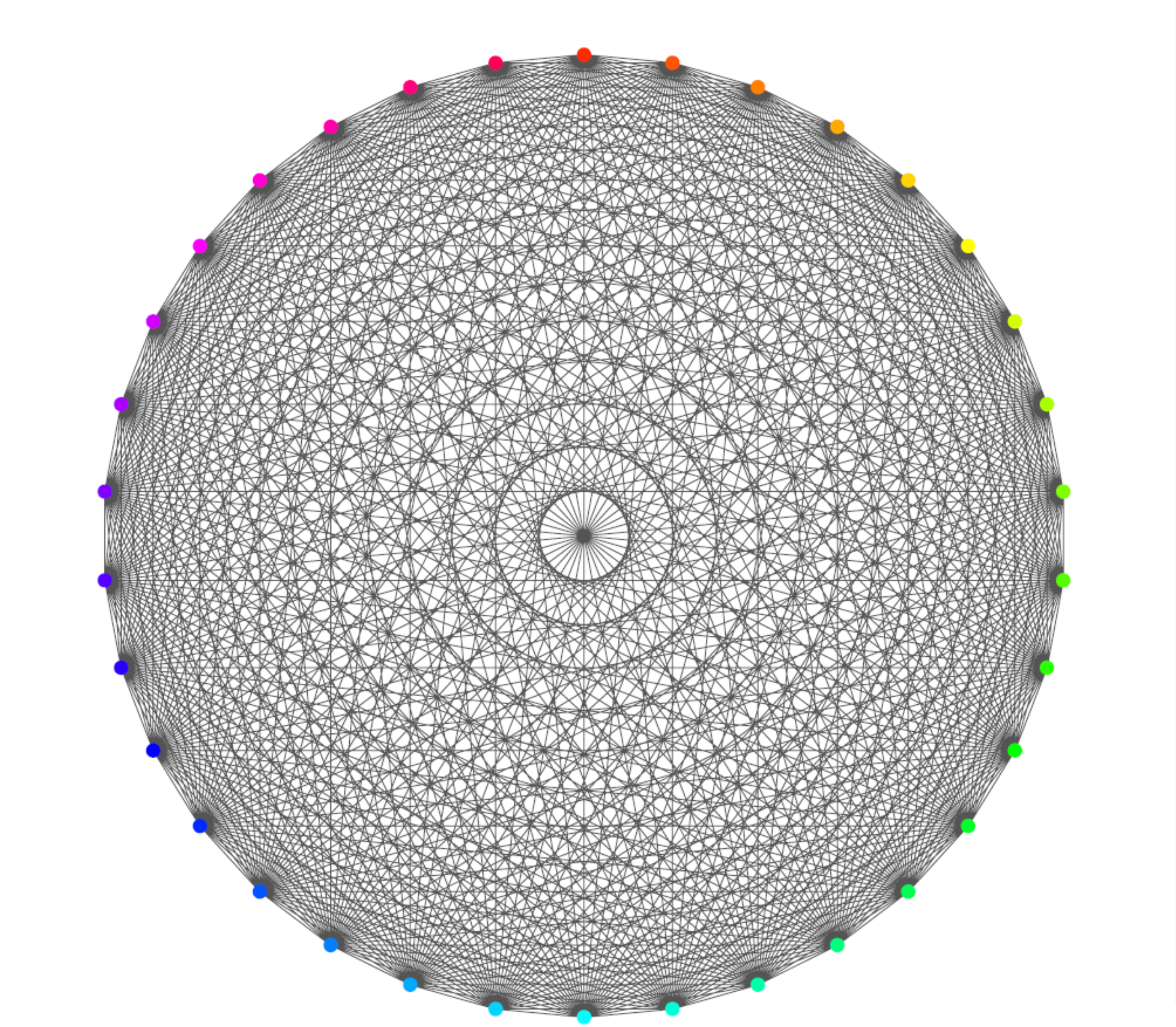
```
Meta23:  (7) [174 0 R]      /T:XObject  /S:Form
Meta24:  (7) [178 0 R]      /T:XObject  /S:Form
Meta25:  (7) [182 0 R]      /T:XObject  /S:Form
Meta46:  (7) [186 0 R]      /T:XObject  /S:Form
Meta47:  (7) [190 0 R]      /T:XObject  /S:Form
Meta48:  (7) [194 0 R]      /T:XObject  /S:Form
Meta49:  (7) [198 0 R]      /T:XObject  /S:Form
Meta50:  (7) [202 0 R]      /T:XObject  /S:Form
Meta51:  (7) [206 0 R]      /T:XObject  /S:Form
Meta68:  (7) [210 0 R]      /T:XObject  /S:Form
Meta69:  (7) [211 0 R]      /T:XObject  /S:Form
Meta70:  (7) [212 0 R]      /T:XObject  /S:Form
Meta71:  (7) [213 0 R]      /T:XObject  /S:Form
Meta72:  (7) [214 0 R]      /T:XObject  /S:Form
Meta73:  (7) [215 0 R]      /T:XObject  /S:Form
Meta74:  (7) [216 0 R]      /T:XObject  /S:Form
Meta75:  (7) [217 0 R]      /T:XObject  /S:Form
Meta76:  (7) [218 0 R]      /T:XObject  /S:Form
Meta77:  (7) [219 0 R]      /T:XObject  /S:Form
Meta78:  (7) [220 0 R]      /T:XObject  /S:Form
Meta79:  (7) [221 0 R]      /T:XObject  /S:Form
Meta80:  (7) [222 0 R]      /T:XObject  /S:Form
Meta81:  (7) [223 0 R]      /T:XObject  /S:Form
Meta82:  (7) [224 0 R]      /T:XObject  /S:Form
Meta83:  (7) [225 0 R]      /T:XObject  /S:Form
Meta84:  (7) [226 0 R]      /T:XObject  /S:Form
Meta85:  (7) [227 0 R]      /T:XObject  /S:Form
Meta86:  (7) [228 0 R]      /T:XObject  /S:Form
Meta87:  (7) [229 0 R]      /T:XObject  /S:Form
Meta88:  (7) [230 0 R]      /T:XObject  /S:Form
Meta89:  (7) [231 0 R]      /T:XObject  /S:Form
Meta90:  (7) [232 0 R]      /T:XObject  /S:Form
Meta91:  (7) [233 0 R]      /T:XObject  /S:Form
Meta94:  (7) [234 0 R]      /T:XObject  /S:Form
```



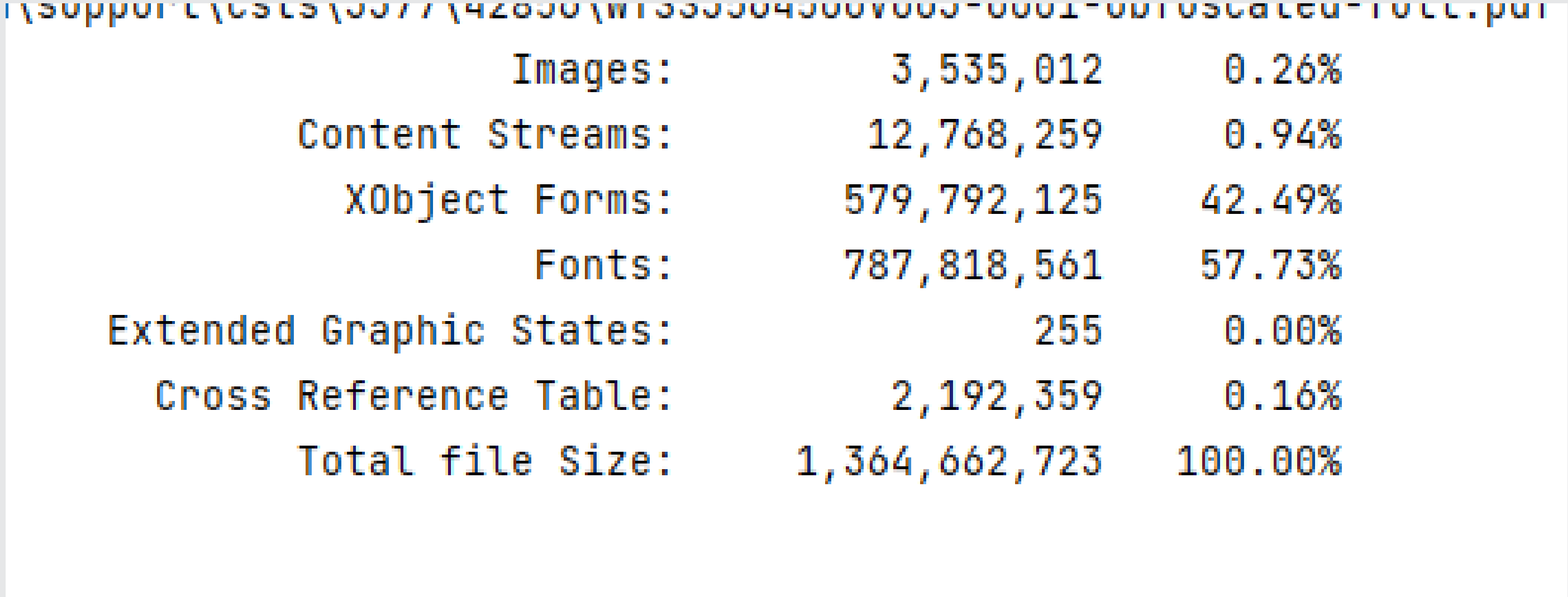
```
Image64:  (10) [167 0 R]      /T:XObject  /S:Image
Image66:  (10) [169 0 R]      /T:XObject  /S:Image
Image9:  (9) [171 0 R]      /T:XObject  /S:Image
Image92:  (10) [172 0 R]      /T:XObject  /S:Image
Meta23:  (7) [174 0 R]      /T:XObject  /S:Form
Meta24:  (7) [178 0 R]      /T:XObject  /S:Form
Meta25:  (7) [182 0 R]      /T:XObject  /S:Form
Meta46:  (7) [186 0 R]      /T:XObject  /S:Form
  BBox:  (4)
  Filter:  FlateDecode
  Length:  1229
  Matrix:  (6)
  Resources:  (4)
    ExtGState:  (2)
    Font:  (4)
    ProcSet:  (5)
    XObject:  (56)
      Image17:  (10) [131 0 R]      /T:XObject  /S:Image
      Image19:  (10) [133 0 R]      /T:XObject  /S:Image
      Image21:  (10) [135 0 R]      /T:XObject  /S:Image
      Image26:  (10) [137 0 R]      /T:XObject  /S:Image
      Image28:  (10) [139 0 R]      /T:XObject  /S:Image
      Image30:  (10) [141 0 R]      /T:XObject  /S:Image
      Image32:  (10) [143 0 R]      /T:XObject  /S:Image
      Image34:  (10) [145 0 R]      /T:XObject  /S:Image
      Image36:  (10) [147 0 R]      /T:XObject  /S:Image
      Image38:  (10) [149 0 R]      /T:XObject  /S:Image
      Image40:  (10) [151 0 R]      /T:XObject  /S:Image
      Image42:  (10) [153 0 R]      /T:XObject  /S:Image
      Image44:  (10) [155 0 R]      /T:XObject  /S:Image
      Image54:  (10) [157 0 R]      /T:XObject  /S:Image
      Image56:  (10) [159 0 R]      /T:XObject  /S:Image
      Image58:  (10) [161 0 R]      /T:XObject  /S:Image
```

Resource Tree Wheel

- Each Form XObject contained an identical **copy** of a previously shared common Resource tree.
- Gordian knot cut by…examining which resources each Form XObject actually used.



42856: Font Optimization



```
\support\csts\3377\42856\W1333504506V003-0001-obfuscated-font.pdf
                    Images:        3,535,012     0.26%
           Content Streams:       12,768,259     0.94%
             XObject Forms:      579,792,125    42.49%
                     Fonts:      787,818,561    57.73%
  Extended Graphic States:               255     0.00%
     Cross Reference Table:        2,192,359     0.16%
           Total file Size:    1,364,662,723   100.00%
```

Font Stats

- **SELECT COUNT(*) from Font where docid=1**
- (416586,)

- **SELECT COUNT(*),subtype from Font where docid=1 and subtype!="Type0" group by subtype**
- (1, 'CIDFontType0')
- (208291, 'CIDFontType2')
- (2, 'Type1')

- **SELECT DISTINCT BASEFONT from Font where docid=1**
- ('Gotham-Bold',)
- ('NimbusSanL-Regu',)
- ('Gotham-Medium',)
- ('Gotham-Book',)
- ('AllAndNone',)
- ('Helvetica',)
- ('ArialMT',)
- ('Gotham-Light',)
- ('DejaVuSans',)

- **SELECT DISTINCT substr(FONTNAME,7) from Fontdescriptor where FONTNAME like "%+%" and docid=1**
- ('+Gotham-Bold',)
- ('+NimbusSanL-Regu',)
- ('+Gotham-Medium',)
- ('+Gotham-Book',)
- ('+Gotham-Light',)
- ('+DejaVuSans',)

- **select count(OBJECTID),basefont from FONT where docid=1 and subtype!="Type0" group by basefont**
- (1, 'AllAndNone')
- (1, 'ArialMT')
- (3244, 'DejaVuSans')
- (50230, 'Gotham-Bold')
- (50586, 'Gotham-Book')
- (3771, 'Gotham-Light')
- (50230, 'Gotham-Medium')
- (1, 'Helvetica')
- (50230, 'NimbusSanL-Regu')

Take-aways

- PDFs are not (always) Trees.
  - Tree nodes aren’t referenced more than once
- PDFs are databases
  - of indirect objects
  - of serialized data structures
- Standard PDF processing can sometimes lead to Horrible Files.
  - But not that often.

Thank you!

working together!

PDF association