

# Born Broken: Fonts And Information Loss In Legacy Documents

Geoffrey Brown and Kam Woods

Indiana University School of Informatics and Computing

# Key Questions

---

- How pervasive are font substitution problems ?
- What information is available to identify fonts ?
- How well can we match the fonts required by a document collection ?
- How can we assist archivists in identifying serious font issues ?

# Font Substitution

---

```
Lbl A:ClrHome
If N<4:Goto Ø
CubicReg L1,L2:3→T:"aX3+bX2+cX+d"→Y1
Disp "aX3+bX2+cX+d"
```

Correctly Rendered

```
Lbl A:ClrHome
If N<4:Goto 0
CubicReg L ,L,:3üT:"aXÓ+bXÜ+cX+d"üY
Disp "aXÓ+bXÜ+cX+d"
```

Default Substitution

## T183plus Font

# Font Substitution (cont.)

---

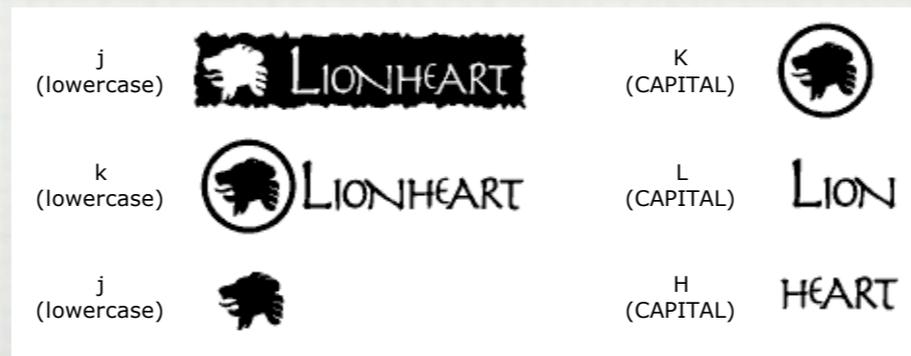
**\*0260931\***

Default Substitution



Forced Substitution of “Code39Azalea”

Barcode 3 of 9 by Request



<http://bowerwebsolutions.com/services/logotruetype.htm>

# Test Collections

---

- ~125,000 Collected Using Glossaries [Reichherzer and Brown 2006]
  - 3910 Total Fonts
  - Top 10 -- Times New Roman, Arial, Times, Verdana, Courier New, Symbol, Tahoma, Arial Narrow, Garamond, Helvetica
  
- ~105,000 Collected From .gov Sites
  - 1920 Total Fonts
  - Top 10 – Times New Roman, Arial, Courier New, Verdana, Times, Arial Narrow, Courier, Tahoma, CG Times, Helvetica

# Font “Collection”

---

Foundry		
Source	Data Type	Number of Fonts
Adobe	Published Table	2374
Bitstream	TrueType Fonts	1556
FontFont	Foundry Supplied Table	11973
URW	Foundry Supplied Table	2358
Operating System		
Microsoft Windows + Office	Font Files	444
Mac OS X + Office	Font Files	322
Application		
Adobe PostScript 3 Fonts	Published List	103
Microsoft Applications	Published List	537
WordPerfect	TrueType Fonts	1080

# Name Problem

Printed by geobrown

Sep 10, 09 10:51 arial.txt Page 1/2

Arial  
Arial  
Arial (PCL6)  
Arial (TT)  
Arial (TT) Bold  
Arial (W1)  
Arial Bold  
Arial Cyr  
Arial Fett  
Arial Greek  
Arial Italic  
Arial MS  
Arial MT  
Arial MT Black  
Arial MT Condensed Light  
Arial MÃ¸ori  
Arial Narrow  
Arial Narrow Bold  
Arial Narrow Italic  
Arial Negrita  
Arial Normaali  
Arial Rounded MT  
Arial Rounded MT Bold  
Arial Unicode MS  
Arial bold  
Arial+0  
Arial, Helvetica, sans-serif  
Arial,Bold  
Arial,BoldItalic  
Arial,Helvetica,sans-serif  
Arial,Italic  
Arial-Black

Sep 10, 09 10:51 arial.txt Page 2/2

Arial-Bold+1  
Arial-BoldItalicMT  
Arial-BoldMS  
Arial-BoldMT  
Arial-ItalicMT  
ArialBlack  
ArialBlack,Italic  
ArialEuroMT  
ArialMS  
ArialMT  
ArialMT-Bold  
ArialMT-ExtraBold  
ArialNarrow  
ArialNarrow,Bold  
ArialNarrow-Bold  
ArialNarrow-BoldItalic  
ArialNarrow-Italic  
ArialUnicodeMS  
BGLLAI+Arial  
BGLLIN+Arial,Bold  
CAPNCH+Arial  
CGIFAD+Arial,BoldItalic  
GPFHEB+Arial  
GPFHJN+Arial,Bold  
H-Arial  
HellasArial  
Verdana, Arial, Times

# Font Information (Truetype)

## Required Tables

Tag	Name
<b>cmap</b>	Character to glyph mapping
<b>head</b>	Font header
<b>hhea</b>	Horizontal header
<b>hmtx</b>	Horizontal metrics
<b>maxp</b>	Maximum profile
<b>name</b>	Naming table
<b>OS/2</b>	OS/2 and Windows specific metrics
<b>post</b>	PostScript information

Each *NameRecord* looks like this:

Type	Name	Description
USHORT	platformID	Platform ID.
USHORT	encodingID	Platform-specific encoding ID.
USHORT	languageID	Language ID.
USHORT	nameID	Name ID.
USHORT	length	String length (in bytes).
USHORT	offset	String offset from start of storage area (in bytes).

BYTE	<b>panose[10]</b>	
ULONG	<b>ulUnicodeRange1</b>	Bits 0-31
ULONG	<b>ulUnicodeRange2</b>	Bits 32-63
ULONG	<b>ulUnicodeRange3</b>	Bits 64-95
ULONG	<b>ulUnicodeRange4</b>	Bits 96-127
CHAR	<b>achVendID[4]</b>	

<http://www.microsoft.com/typography/otspec/>

# Word Document Font Information

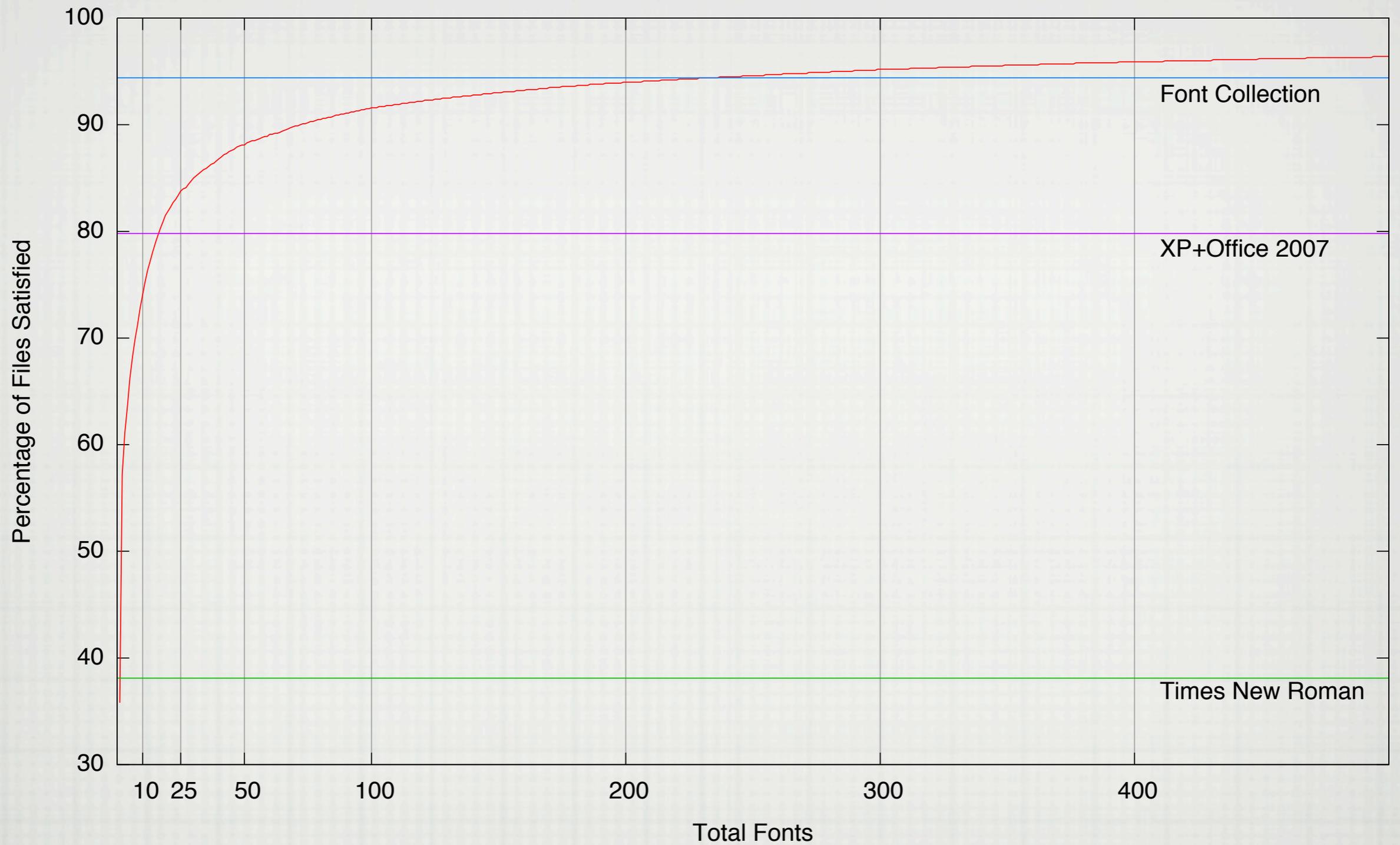
## Font Family Name (FFN)

<b>b<sub>10</sub></b>	<b>b<sub>16</sub></b>	<b>Field</b>	<b>Type</b>	<b>Size</b>	<b>Bitfield</b>	<b>Comment</b>
0	0	cbFfnM1	uns char			Total length of FFN - 1.
1	1	prq	uns char	:2	03	Pitch request
		fTrueType	uns char	:1	04	When 1, font is a TrueType font
			uns char	:1	08	Reserved
		ff	uns char	:3	70	Font family id
			uns char	:1	80	Reserved
2	2	wWeight	short			Base weight of font
4	4	chs	uns char			Character set identifier
5	5	ixchSzAlt	uns char			Index into <code>ffn.szFfn</code> to the name of the alternate font
6	6	panose	PANOSE			
16	10	fs	FONTSIGNATURE			
40	28	xsZFfn	XCHAR[]			Zero terminated string that records name of font. Possibly followed by a second <code>xsZ</code> which records the name of an alternate font to use if the first named font does not exist on this system. Maximal size of <code>xsZFfn</code> is 65 characters.

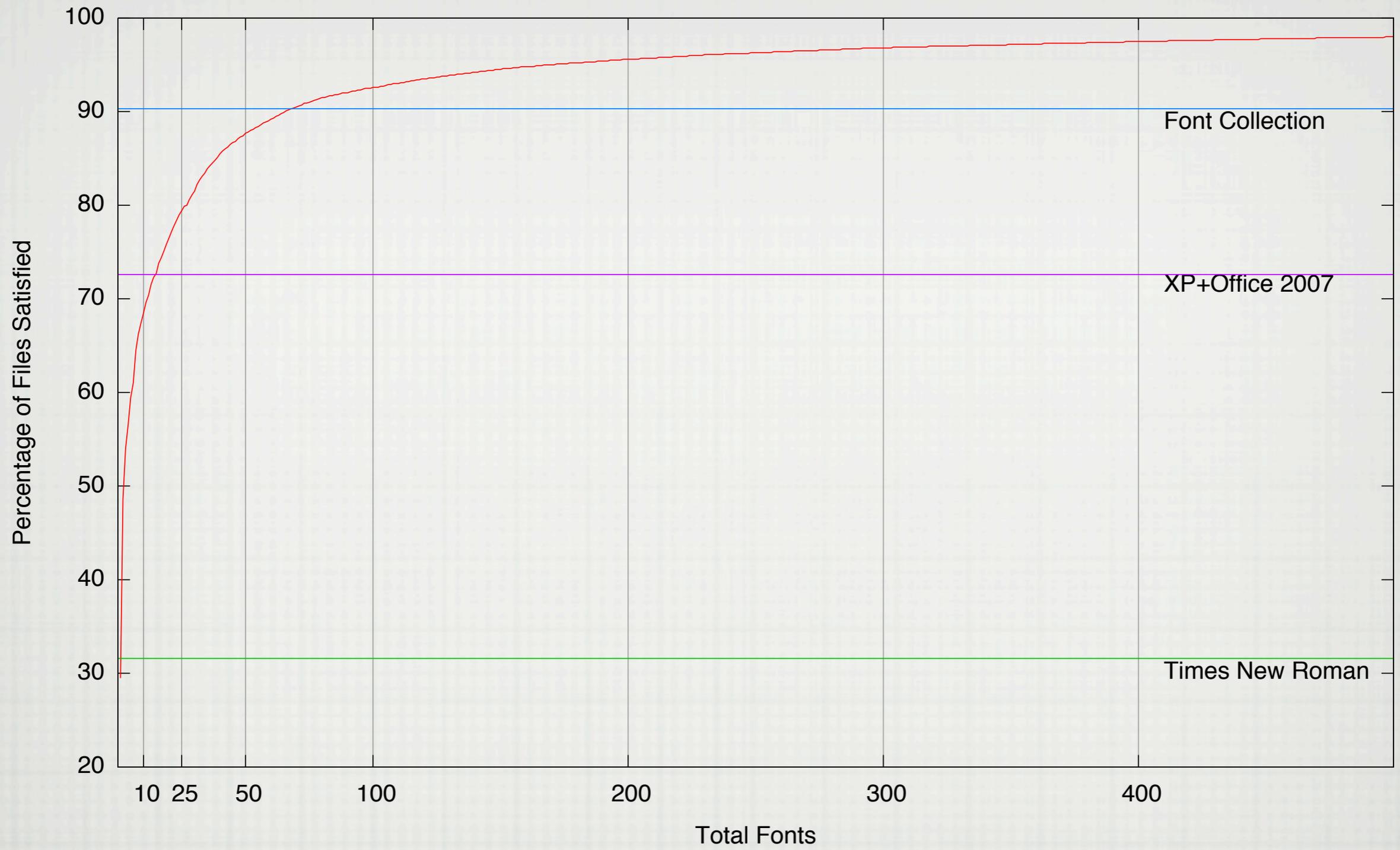
# Extracting Fonts from Documents

---

- libwv (used in abiword) + bug fixes on font identification
- Custom C program (several hundred lines)
- Walk document character-by-character
- Process results with shell scripts
- Match name strings (complete match) against font database



Documents collected using Glossaries



.gov Documents

# Windows Font Analysis Tool

The screenshot displays a Windows desktop environment. In the foreground, a window titled "Font Analysis" is open, with a "Browse For Folder" dialog box overlaid on it. The dialog shows a list of folders, with "WordSampleDocs" selected. To the right, a Notepad window titled "fontinfo.xml" displays the following XML content:

```
<!--FontAnalysis-->
<FontAnalysis xmlns:font="urn:FontAnalysis">
  <file font:file_name="C:\Documents and Settings\iucs\Desktop\DocSmall\AGREEMENT.doc">
    <font font:name="Times New Roman">
      <char>
        <count>622</count>
      </char>
      <glyph>
        <count>52</count>
      </glyph>
    </font>
    <font font:name="Arial Black">
      <char>
        <count>28</count>
      </char>
      <glyph>
        <count>15</count>
      </glyph>
    </font>
  </file>
  <file font:file_name="C:\Documents and Settings\iucs\Desktop\DocSmall\Concannon.doc">
    <font font:name="Arial">
      <char>
        <count>3</count>
      </char>
      <glyph>
        <count>2</count>
      </glyph>
    </font>
    <font font:name="Times New Roman">
      <char>
        <count>18639</count>
      </char>
      <glyph>
        <count>81</count>
      </glyph>
    </font>
    <font font:name="symbol">
      <char>
        <count>2</count>
      </char>
    </font>
  </file>
</FontAnalysis>
```

The taskbar at the bottom shows the Start button, several open applications including "DocAnalyzerInterfac...", "Font Analysis", and "fontinfo.xml - Notepad", and the system tray with the time "1:35 PM".

# Word Add-In

The screenshot shows the Microsoft Word interface with the 'Add-Ins' ribbon selected. Two add-in windows are open: 'Font Analyzer' and 'Search for Font Occurrence'. The 'Font Analyzer' window displays the following information:

- Number of characters: 4577
- Dump of font name table:
  - Name: Times New Roman Number of chars: 4567 Number of glyphs: 69
  - Name: Barcode 3 of 9 by request Number of chars: 10 Number of glyphs: 8

The 'Search for Font Occurrence' window shows the following options:

- Times New Roman
- Barcode 3 of 9 by request

Buttons for 'Next Occurance' and 'Close' are visible in this window.

The main document content includes:

East Asia has been quite successful in reducing the proportion of people who suffer from **hunger**, while Africa's malnutrition rate has hardly budged.

**2. Paragraph 111**

The last sentence *should read*:

This is why I intend to submit to the General Assembly in September 2002 a report that will propose further programmatic, institutional and process improvements, so that we can translate the ambitious template of the Declaration into an achievable agenda of action.

**3. Annex table, goal 6**

Replace the text of goal 6 by the following:

02-60931 (E) 300902  
**\*0260931\***

The status bar at the bottom indicates 'Page: 1 of 5', 'Words: 646', and '110%' zoom. The taskbar shows the Start button and several open applications, including WordDocAnalyzerAd..., death\_rate29.doc [C...], Font Analyzer, and Search for Font Occu...

# Logo Example

The screenshot shows a Microsoft Word window titled "deflation68.doc [Compatibility Mode] - Microsoft Word". The ribbon includes Home, Insert, Page Layout, References, Mailings, Review, View, Developer, and Add-Ins. A "Scan Document" task pane is on the left. The main document area contains a logo with the Greek characters "εφγ ιφ" in a blue box, followed by the text "AMP4" and "MONITORING PLAN". Below this is a paragraph of text: "Company strategy for 2005-2010 - its core environmental improvements, services to customers and prices." Two windows are open: "Font Analyzer" and "Search for Font Occurance".

**Font Analyzer** window content:

```
Number of characters: 56085  
Dump of font name table:  
Name: OFWAT Logo Number of chars: 6 Number of glyphs: 6  
Name: Arial Number of chars: 56079 Number of glyphs: 83
```

**Search for Font Occurance** window content:

- OFWAT Logo
- Arial

Buttons: Next Occurance, Close

Taskbar: Page: 1 of 1, Words: 8,223, WordDocAnalyzerAd..., deflation68.doc [Com..., Font Analyzer, Search for Font Occu..., 75%, 10:34 AM

# Cyrillic Font

The screenshot shows a Microsoft Word document with the following elements:

- Search for Font Occurance dialog:** A modal dialog box with radio buttons for font selection: CyrillicLaser, Times New Roman, GlasnostLight (selected), and Helvetica. It includes "Next Occurance" and "Close" buttons.
- Font Analyzer dialog:** A dialog box displaying font statistics: "Number of characters: 1617". It lists a "Dump of font name table" with three entries: CyrillicLaser (102 chars, 6 glyphs), Times New Roman (57 chars, 27 glyphs), and Helvetica (96 chars, 30 glyphs). It includes "Search Document", "Markup Document", and "Close" buttons.
- Document Content:** The document features the Washington State Department of Transportation logo and the title "TITLE VI PUBLIC INVOLVEMENT". The main text is in a monospaced font and contains Cyrillic characters, such as "Hfpltk VI Pfrjyf j uhf;lfycrb[ ghfdf[ 1964 ujlf j,z eghfdktybt infnf Dfibyyny (Washington State Department ufhfynbhjdfnm> xnj dct ;bntkb hfgjyid> ult yfvtxf/,elen bvtnm djpvi;yjcnm dscrpfnm cdjt vytybt gj gidile nhfycgjhnysh[ ghjuhfvv b vthjghbznbg> rjnjhst vjuen rjcyenmcz b[ hfgjyf. Xnj,s givixm d 'njv> vs ghjcbv dfc lj,hjdikmyj ghtlicnfdbnm yfv byajhvfwb/ j dfitg hfct> 'nybxteriq ghbyflkt;yjcnb b#bkb gikt. Ds yt j,zpfys ghtlicnfdbnm pfgjhityye/ byajhvfwb/> xnj,s ghbyznm exfcnbt d 'njv cj,hfybb. Rjhlbyfnjh ckeifybq WSDOT b Jabc hfdys[ djpvi;yjcntq (Office of Equal Opportunity, CEO) ...".
- Word Ribbon:** The "Review" tab is active, showing options for "Scan Document" and "Font Analyzer".
- Taskbar:** Shows the Start button and several open applications: WordDocAnalyzerAd..., RussianPubInvolv.do..., Font Analyzer, Search for Font Occu..., and Firefox Updated - Mo... The system clock shows 11:03 AM.

# Problems

---

- Old document (pre Unicode) used ascii character range for Cyrillic.
- Modern versions of “Glasnost Light” use unicode pages properly. Old font is not available  
It is with profound regret that we inform you that Casady & Greene will close its doors on July 3rd, 2003,
- Legacy substitution documentation identified Corel font “Czar”, but that’s not quite right  
[http://nwalsh.com/comp.fonts/FAQ/cf\\_33.htm](http://nwalsh.com/comp.fonts/FAQ/cf_33.htm)
- The Cyrillic Charset Soup -- <http://czyborra.com/charsets/cyrillic.html>

## WD2000: Incorrect Characters Appear When You Open Document in Earlier Eastern European Version of Word

This article was previously published under Q260162

### SYMPTOMS

---

When you open a Word document that was created using an earlier Eastern European version of Microsoft Word, the text of the document may appear as square boxes or other incorrect characters when you open the document in the English (U.S.) version of Microsoft Word.

### CAUSE

---

This problem is caused by font-mapping and character-mapping problems. For example, this problem can occur when you create a document by using a font that is not installed on your computer.

### RESOLUTION

---

To correct these problems, Microsoft provides a Word Font Repair Macro that converts the fonts in the document (or the selected text) to Arial.

The following file is available for download from the Microsoft Download Center:

[Eefonts.exe](http://download.microsoft.com/download/word2000/eefonts/2000/w9xnt4/en-us/eefonts.exe) (http://download.microsoft.com/download/word2000/eefonts/2000/w9xnt4/en-us/eefonts.exe)      Release Date: 27-Jan-2000

For additional information about how to download Microsoft Support files, click the following article number to view the article in the Microsoft Knowledge Base: [119591](http://support.microsoft.com/kb/119591/EN-US/) (http://support.microsoft.com/kb/119591/EN-US/ ) How to Obtain Microsoft Support Files from Online Services      Microsoft scanned this file for viruses. Microsoft used the most current virus-detection software that was available on the date that the file was posted. The file is stored on security-enhanced servers that help to prevent any unauthorized changes to the file.

### MORE INFORMATION

---

To download and install the Word Font Repair Macro, follow these steps:

1. Click the Eefonts.exe link in the "Resolution" section of this article to download the Eefonts.exe file.
2. In the **File Download** dialog box, click to select **Save this program to disk**, and then click **OK**.
3. Change the **Save in** box to the folder where you want to save the Eefonts.exe file, and then click **Save**.
4. After the file has been downloaded, install the macro. To do this, follow these steps:

a. Quit Microsoft Word 2000.

# What can we do ?

---

- A better job of identifying “critical” font substitutions
  - Use unicode range to assist language identification
  - Use language identification tools to identify human language vs symbolic font use
- Create better tools to facilitate quality assurance testing by isolating text with font substitutions
- Create a clearinghouse of font identification and substitution information