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

TI83plus 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		
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

Source	Data Type	Number of Fonts

Name Problem

Printed by geobrown


```
 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
```



```
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*
cmap	Character to glyph mapping
head	Font header
hhea	Horizontal header
hmtx	Horizontal metrics
maxp	Maximum profile
name	Naming table
OS/2	OS/2 and Windows specific metrics
post	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	panose[10]	
ULONG	ulUnicodeRange1	Bits 0-31
ULONG	ulUnicodeRange2	Bits 32-63
ULONG	ulUnicodeRange3	Bits 64-95
ULONG	ulUnicodeRange4	Bits 96-127
CHAR	achVendID[4]	

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

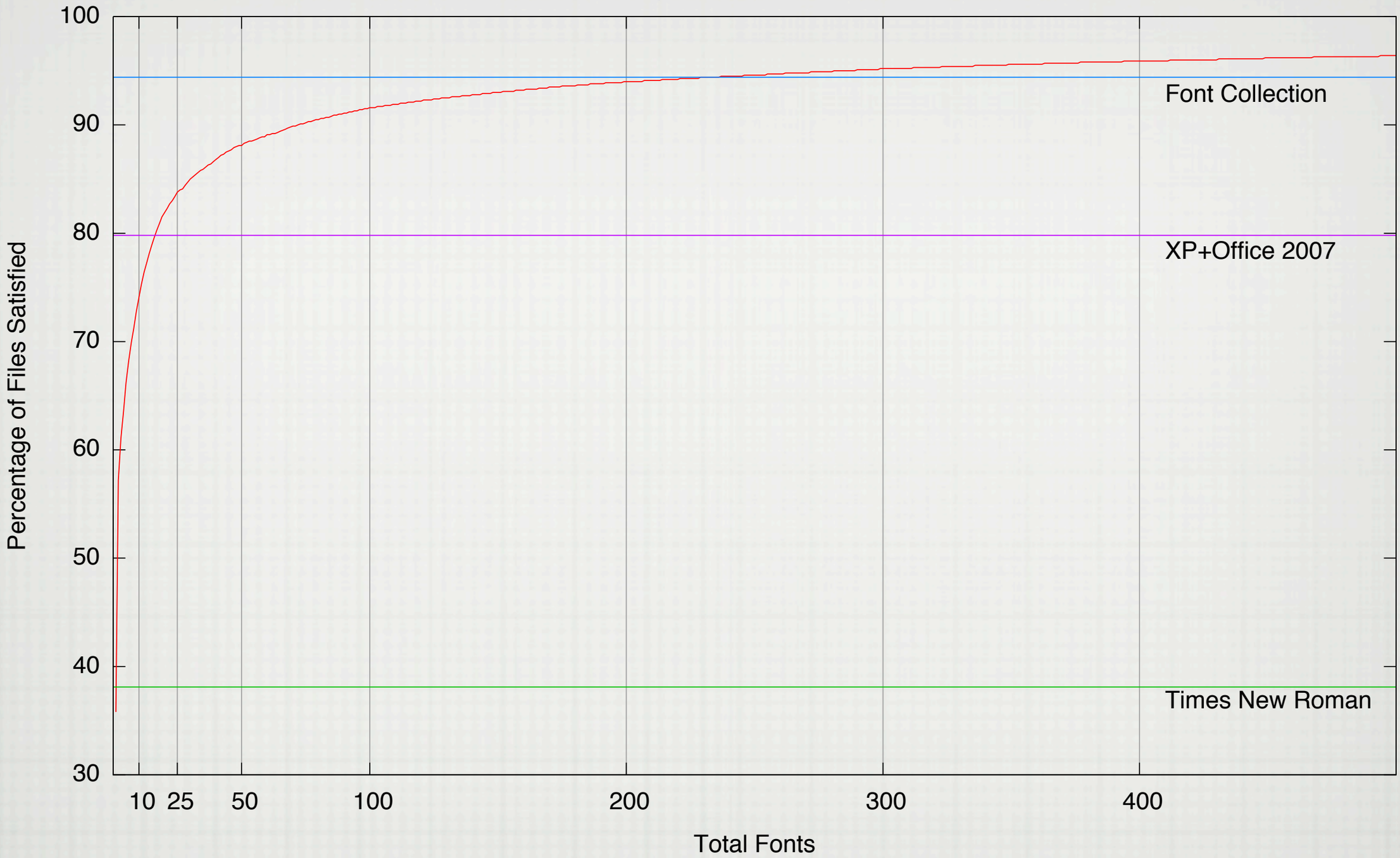
Word Document Font Information

Font Family Name (FFN)

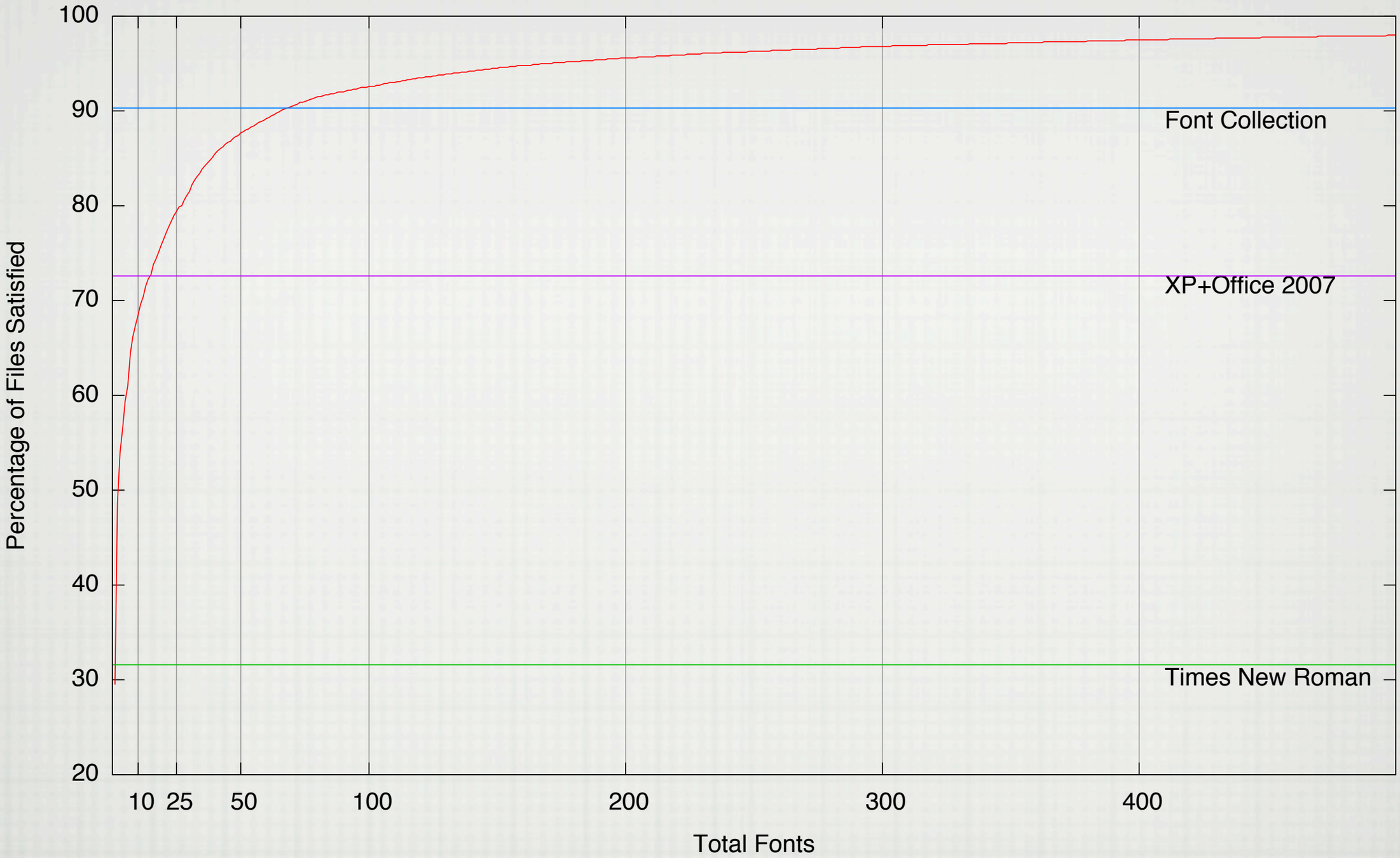
b_{10}	b_{16}	Field	Type	Size	Bitfield	Comment
0	0	`cbFfnM1`	uns char			Total length of FFN - 1.
1	1	`prq`	uns char	:2	03	Pitch request
		`fTrueType`	uns char	:1	04	When 1, font is a TrueType font
			uns char	:1	08	Reserved
		`ff`	uns char	:3	70	Font family id
			uns char	:1	80	Reserved
2	2	`wWeight`	short			Base weight of font
4	4	`chs`	uns char			Character set identifier
5	5	`ixchSzAlt`	uns char			Index into `ffn.szFfn` 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 `xsz` which records the name of an alternate font to use if the first named font does not exist on this system. Maximal size of `xszFfn` 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

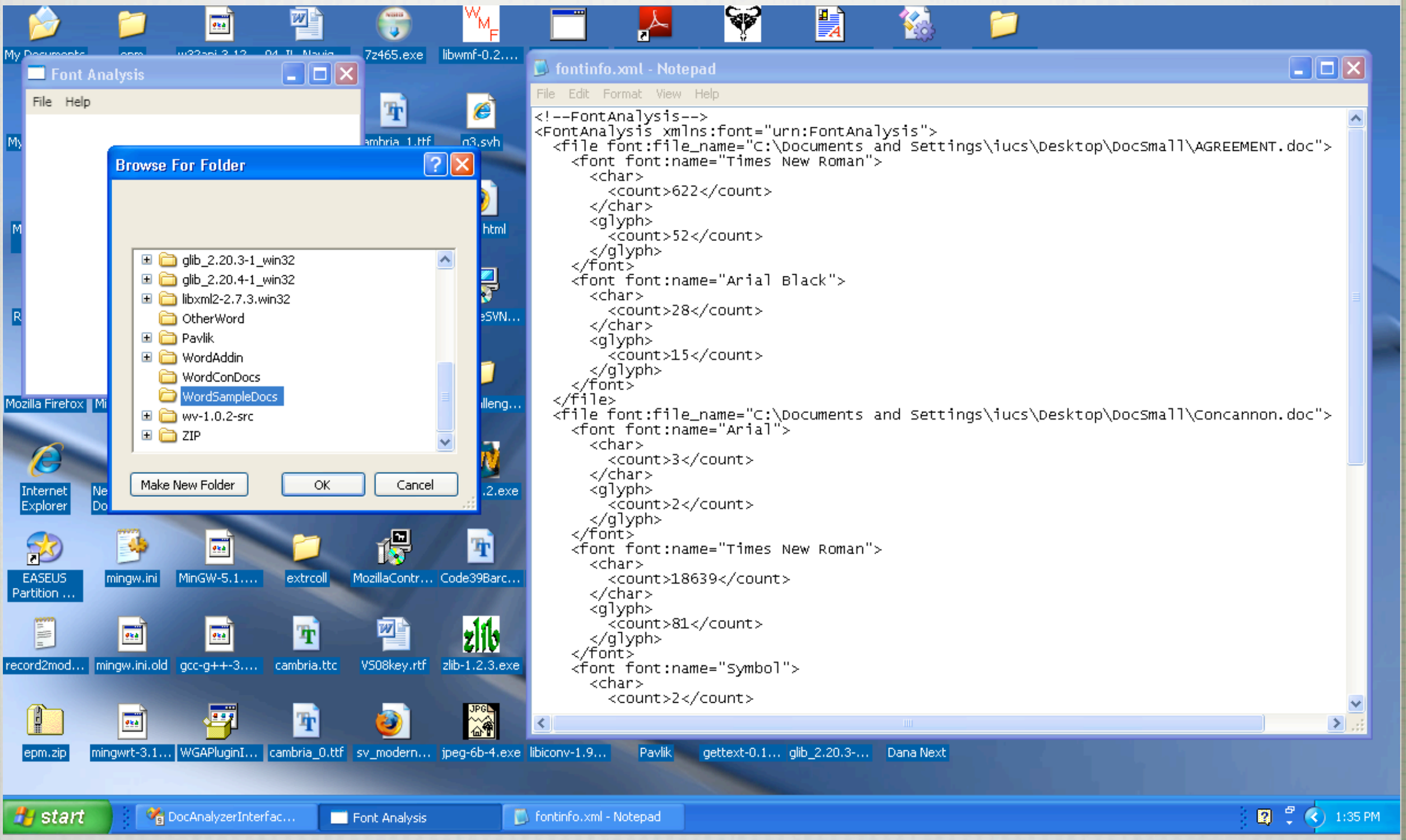


Documents collected using Glossaries

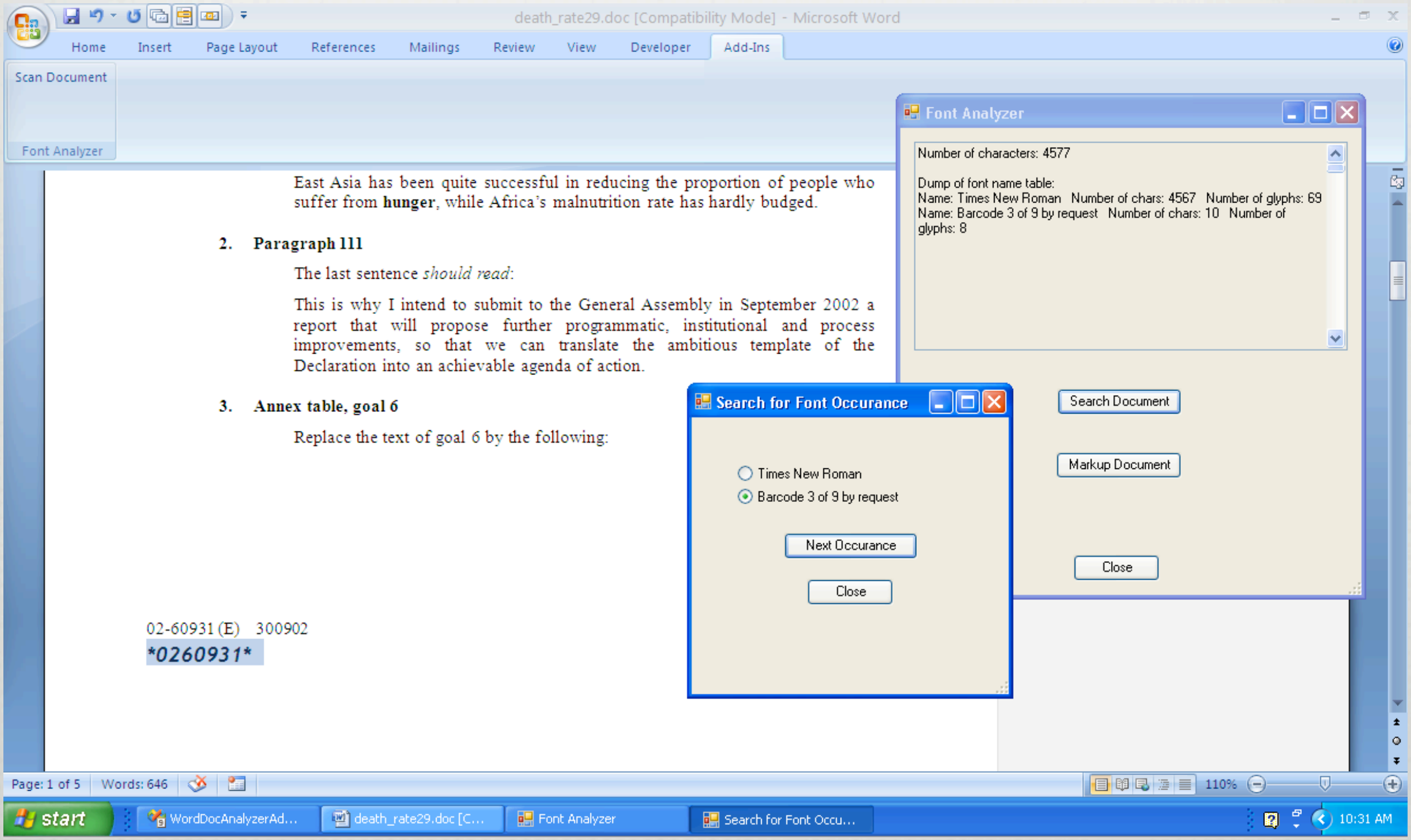


.gov Documents

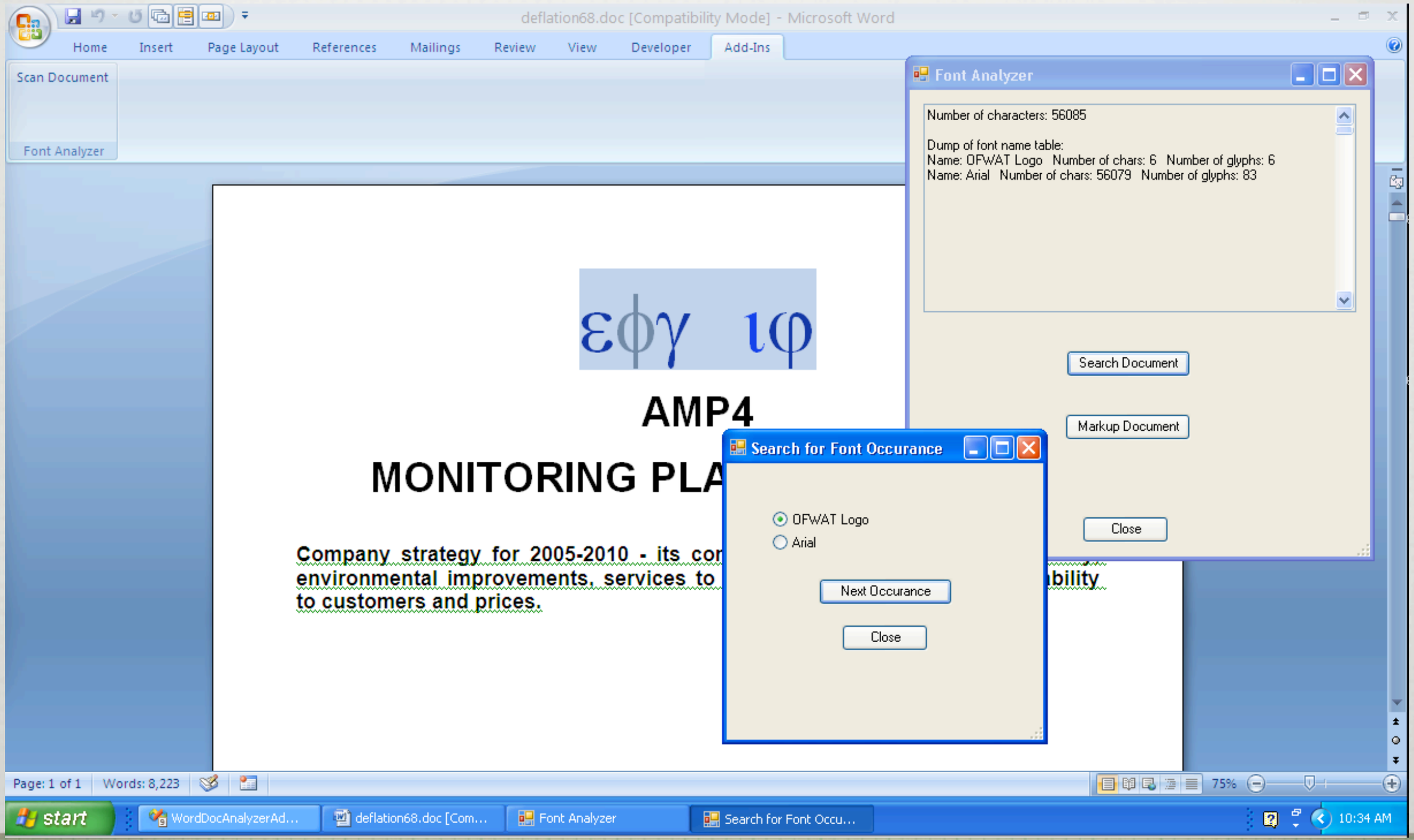
Windows Font Analysis Tool



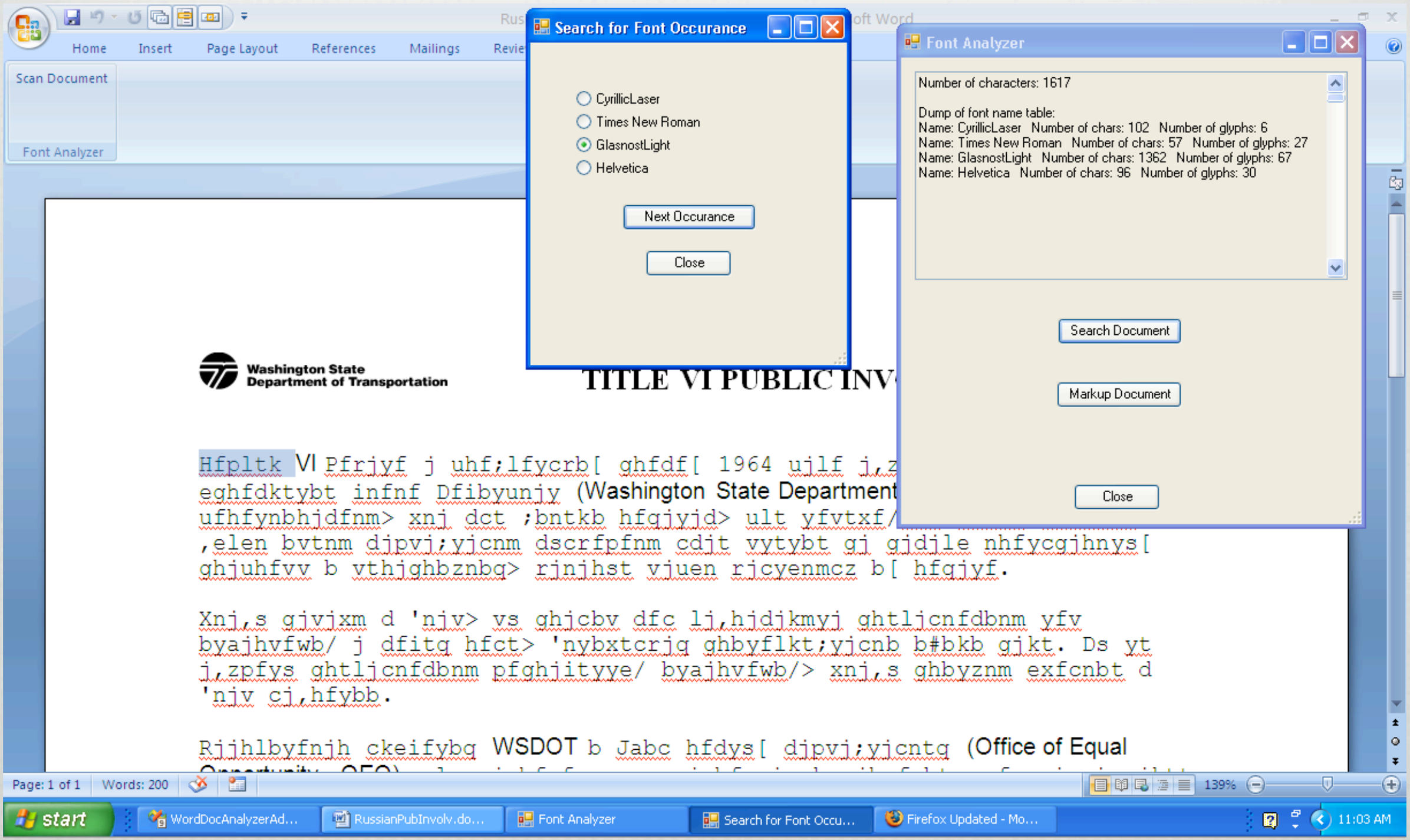
Word Add-In


death_rate29.doc [Compatibility Mode] - Microsoft Word
Home
Insert
Page Layout
References
Mailings
Review
View
Developer
Add-Ins
Scan Document
Font Analyzer
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
Font Analyzer
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
Search Document
Markup Document
Close
Search for Font Occurance
Times New Roman
Barcode 3 of 9 by request
Next Occurance
Close
Page: 1 of 5
Words: 646
110%
start
WordDocAnalyzerAd...
death_rate29.doc [C...
Font Analyzer
Search for Font Occu...
10:31 AM

Logo Example



Cyrillic Font



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
- The Cyrillic Charset Soup -- http://czyborra.com/charsets/cyrillic.html

Article ID: 260162 - Last Review: August 5, 2004 - Revision: 2.2

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) 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/) 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