ISO/IEC JTC1/SC2/WG2/IRG N1702

Universal Multiple-Octet Coded Character Set
International Organization for Standardization

Doc Type: ISO/IEC JTC1/SC2/WG2/IRG

Title: Font Encoding Suggestions for Multi-column Code Charts

Source: Dr. Ken Lunde, Adobe Systems Incorporated (lunde@adobe.com)

Status: Individual Contribution

Action: For consideration by the IRG

Date: 2010-06-24

Background

In order to expedite the editing of the multi-column CJK Unified Ideograph code charts by the ISO 10646 editor, the fonts that are supplied by each National Body should ideally encode their glyphs according to their corresponding CJK Unified Ideograph code points. Doing so significantly minimizes the possibility of error.

The mapping from code points to glyphs is encapsulated in only one font table, specifically the 'cmap' table. There are freely-available and easy-to-use font tools that simplify the modification of the mappings in the 'cmap' table. This document provides suggestions, guidance, and best practices for preparing fonts for use in the multi-column code charts.

Recommended Font Tools

The easiest font tool for re-encoding an sfnt-based font, meaning a TrueType or OpenType, is called *ttx* (aka, *TTX/FontTools*). It is a cross-platform command-line tool, and can be downloaded from the following URL:

http://sourceforge.net/projects/fonttools/

Its operation is simple. When specifying a font as input, it outputs an XML file that uses "ttx" as its filename extension:

```
% ttx font.ttf
```

The output file from the above command line is thus *font.ttx*.

For BMP code points, a Format 4 'cmap' table section must be included, and uses the following tags:

```
<cmap_format_4 platformID="3" platEncID="1" language="0">

</cmap_format_4>
```

The values for the "platformID," "platEncID," and "language" attributes should be specified as shown above, specifically 3, 1, and 0, respectively.

For code points beyond the BMP, a Format 12 'cmap' table section must be included, or added if the original font did not include one. The following tags must be used:

```
<cmap_format_12 platformID="3" platEncID="10" format="12" reserved="0" length="0" language="0" nGroups="0">

</cmap_format_12>
```

Note that zero (0) should be used for the values of the "length" and "nGroups" attributes, because they will be calculated anyway by *ttx* when the font is recompiled. The vales for the "platformID," "platEncID," "language," and "format" attributes should be specified as shown above, specifically 3, 10, 0, and 12, respectively.

The character code to glyph mappings that appear between the tags follow a very simple tag format. The following is an example of a BMP (U+4E00) and a non-BMP (U+20000) mapping:

```
<map code="0x4e00" name="glyph_name_1"/>
<map code="0x20000" name="glyph_name_2"/>
```

The Format 4 'cmap' table section can contain only BMP mappings, meaning up to four hexadecimal digits. The Format 12 'cmap' table section includes *both* BMP and non-BMP mappings. The Format 12 'cmap' table is necessary *only* if the font includes non-BMP mappings. For such fonts, the Format 4 'cmap' table section must be a BMP-only subset of the Format 12 'cmap' subtable section. In other words, only the first line above can appear in a Format 4 'cmap' table section, but both should appear in a Format 12 one.

When changing the "code" attribute, it is important to leave the "name" attribute as-is. If you modify the "name" attribute, you will need to also visit the 'glyf' or 'CFF' table section to make the same change to the appropriate glyph names. Glyph names are unimportant in this context, so leaving them as-is is best.

When the 'cmap' table section is finished, you recompile the font by simply specifying the XML file as input, as follows:

```
% ttx font.ttx
```

The output will be *font.ttf* or *font.otf*, depending on whether the input file was TrueType or OpenType.

If you have tables that correlate the original code points in the 'cmap' table with the actual CJK Unified Ideograph code points, simple scripting languages can be used to minimize the effort in making the changes as described in this section.

Optional Font Tools

If there is any fear that *ttx* modifies any other tables of the font, you can use the *sfntedit* tool that is included in AFDKO (Adobe Font Development Kit for OpenType) to splice the modified 'cmap' table from the modified font to the original font. AFDKO is available for Windows and Mac OS X, and can be downloaded from the following URL:

http://www.adobe.com/devnet/opentype/afdko/

You should first make a copy of the original font. Then, use *sfntedit*'s "-x" option to extract the 'cmap' table from the font (called *font_new.ttf*) to a file called *cmap.data*, as follows:

```
% sfntedit -x cmap=cmap.data font_new.ttf
```

Then, use *sfntedit*'s "-a" option to splice into the original font (called *font.ttf*) the extracted 'cmap' table (called *cmap.data*), as follows:

```
% sfntedit -a cmap=cmap.data font.ttf
```

It is really that simple.

Other Font Tools

Another useful font tool for tweaking font tables is called *DTL OTMaster*. The "Light" version is free. It is available from the following URL:

http://www.fontmaster.nl/english/OTMaster.html

Further Assistance

I am willing to provide assistance to IRG members for modifying the fonts that are submitted to the ISO 10646 editor, as a way to expedite the process, and to minimize the possibility of errors.

That is all.