Doc Number: L2/02-129

To: **UTC**
Re: **L2/02-127 "Machine-Readable StandardizedVariants.txt" (M. Davis)**
From: **Ken Whistler**
Date: **2002-03-28**

Warning! Warning! Process complexification safety limited exceeded.

What Mark is suggesting here is to take the new file we just created, StandardizedVariants.html, and turn it on its head by creating a separate, machine-readable text file, and using that to generate StandardizedVariants.html. (And doing that for each release of Unicode that touches anything to do with the standardized variant sequences.)

The new file would then be the source file, and StandardizedVariants.html would merely be a derived file.

While this at first all sounds fine and good, there is devious complexity hiding here.

Asmus said:

"The names of the gifs are all derived from their code points, that helps."

Well, that sounds good, and is even true in a way, but isn't complete. The representative glyphs extracted from the code charts do have simple, predictable names:

```
U2229.gif
U1880.gif
```

But the math variant sequence glyphs aren't even gifs:

```
U2225-FE00.jpg
```

And the terminology for them doesn't match the gifs for Mongolian:

```
U1820180Bfina.gif
U1820180Bisol.gif
U1820180Bmedi.gif
```

So either

a) the generating script has to be complexified to special case the name generation by group, or,

b) the names and types of the images needs to be regularized, or,

c) the names and types of the images need to be added to the machine-readable source file as fields.

Choice a) would be ad-hoc hacking, and would eventually cause maintenance problems. Choice b) would force trouble in fixing some part of the large number of images already released, and would cause versioning issues for the next release. (There are, by the way, already 184 individual gifs or jpgs now officially under source control as part of the Unicode Character Database -- a number of individual items that easily swamps all the other files being maintained as part of the UCD, so this is not a trivial change control issue.) Choice c) I see as the only viable, maintainable way to do this that doesn't introduce too large a risk for error introduction and delays for the next release.

Unfortunately, choice c) is almost guaranteed to induce pissing and moaning in whoever has to edit the initial source data file.

Mark said:

> "I generate all the derived data files anyway, I can add a module to generate the HTML without changing our process to add a Perl script."

I agree that this would be preferable to introducing the building, debugging and maintaining of a Perl script for this, when we don't already do that as part of the Unicode UCD update releases.

However, every additional module added to Mark's derivation program (which itself is a gigantic and basically undocumented hack -- although an elegant hack, no doubt), increases our exposure, and the dependency of the entire release process on the time availability and attention of a single person, who himself is generally very overbooked.

And eventually Mark is going to get tired of this (which keeps getting significantly more complicated for each release), and bequeath the whole thing on some player to be named later, and the process will stop dead for awhile until somebody figures out how to make it work again.

Now I'm not saying this is impossible, or even necessarily a bad idea. But it is my duty as the designated naysayer for the UTC to point out the trouble we're getting into if we take this course.

Frankly, I think creating StandardizedVariants***.html*** as our source file for this in the UCD in the first place was a bad idea. I'd prefer a solution that simply has the machine-readable file, including URL's to the appropriate list of change-controlled images. Skip the generation of StandardizedVariants.html as an official file altogether, and let people push the machine-readable file through whatever display renderer or text-fancifier they want, to get nice readable versions of the document. That is more in keeping with our approach to the rest of the data files in the UCD -- and I don't see why we should force ourselves into a difficult position of maintaining custom file transducers as part of the UCD release process so that we can keep on presenting StandardizedVariants.html as a preformatted file for people.

Alternatively, we could distribute a small piece of special-purpose Java or C code that extracts the tables with known, fixed headers rows from StandardizedVariants.html and dumps all the remaining rows of those tables in a machine-readable format. That would seem to be to be an easier solution for the problem that Mark mentioned of not having a machine-readable form of the data. And it would certainly involve less complexification of release process for an already overworked editorial committee.

--Ken