L2/10-458

2010-11-02

Universal Multiple-Octet Coded Character Set
International Organization for Standardization

Doc Type: Working Group Document

Title: Variation Sequence for Emoji

Source: Dr. Ken Lunde & Sairus Patel, Adobe Systems Incorporated

Status: Corporate Full Member Contribution

Action: For consideration by the UTC

Date: 2010-11-02

Background

The first version of Unicode to provide adequate support for *emoji* (aka, emoticons) is Version 6.0, in terms of encoding the characters. Of course, the direct encoding characters is the first step toward making them available for practical use, which in this case clearly refers to font-based implementations.

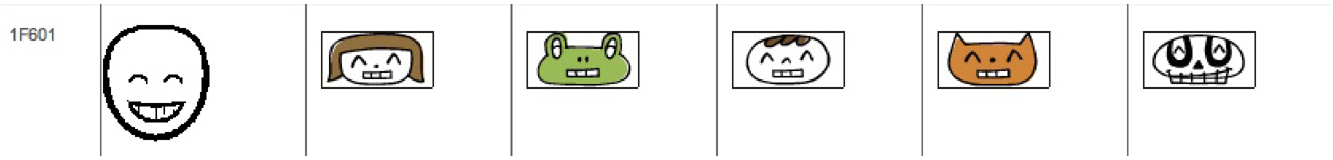
Emoji are unique in that they are often colored or animated, but that's not important in the context of this document. What is important is that they flourish in "plain text" environments, such as in text messages. This document proposes the use of Unicode Variation Sequences (UVSes) as a way to standardize the recording of subtle differences in emoji, whether to preserve distinctions for the benefit of the different carriers, or to support multiple variations within a single font.

Proposal Details

Text messaging is a "plain text" environment, whether the client is a mobile device, or a service such as Twitter. The amount of data that is transmitted must be minimal. Specifically, fonts, font-like data, or multimedia objects cannot be embedded. If it is not "plain text," it does not survive. (To some extent, this data-brevity requirement may have given birth to emoji.)

In terms of valid and real-world "use cases" for variant forms of emoji, we can think of at least two. One is for the carrier-specific forms of emoji that were necessarily unified, and whose distinctions, at least for some of them, have been partially lost as a result. The other is to support additional graphical attributes that are orthogonal to carrier-specific differences.

To describe the latter, some emoji can be enhanced or modified to include additional graphical attributes, above and beyond what has been encoded, yet still convey the same semantics. Many of the emoji in the U+1F6*xx* range fall into this category. Below are example variations for U+1F601 (*GRINNING FACE WITH SMILING EYES*) that include attributes for *girl*, *frog*, *boy*, *cat*, and *skeleton*, yet arguably share the same or similar "typeface design" characteristics:



(As a side note, U+1F601 with the "cat" attribute is seemingly identical to U+1F638 (*GRINNING CAT FACE WITH SMILING EYES*). In looking at the *EmojiSources.txt* file, it seems that this character was included for compatibility purposes, specifically for KDDI's emoji character set.)

We can think of at least three ways in which variant forms of emoji can be implemented as fonts, enumerated as follows:

1. As separate font resources
2. As a single font resource, using OpenType GSUB (*Glyph SUBstitution*) features to access the variant forms
3. In a single font resource, using UVSes to specify variant forms

In a “plain text” environment, only the third solution allows a single font resouce to easily serve the correct glyphs. Furthermore, the UVS mechanism specifies that a reasonable default form will be displayed if the specific glyph that corresponds to the UVS is not included in the selected font.

Consider that even if font developers give away their emoji fonts at no charge, a text message recipient might not have it installed. Also consider that variant forms of an emoji, such as a smiling girl or smiling skeleton, add personalized meaning that is sufficiently important to preserve—when possible, of course, given the constraints of some text messaging protocols—rather than displaying a generic emoji form.

Therefore, transmitting a variant form of an emoji in “plain text” as a UVS would allow for maximum possibility of both sender and recipient seeing the same form. If the recipient’s device does not have a font that includes a glyph for the specified UVS, or simply doesn’t support UVSes, then the emoji that corresponds to the Base Character of the UVS will be displayed, meaning that a user will see the generic emoji for the specified code point, and the basic meaning is thus conveyed.

In closing, we are aware of standardized UVSes, such as for math and Phags-pa, and those that are registered, such as Ideographic Variation Sequences (IVSes). While we feel that UVSes for emoji should be registered, we defer to the UTC for their judgment and expertise.

That is all.