L2/18-044

Re: Encoding emoji vs other characters
From: Mark Davis, Jeremy Burge, Peter Edberg
Date: 2018-01-22

The following discussion was requested by the UTC in action 153-A9 ("Compose text clarifying the different kinds of pictographic characters and the needed justification for encoding as new characters").

This is a rough draft of such text for discussion and revision as necessary during the UTC meeting.

Emoji are very different than other Unicode characters. For many years, they were considered out of scope. Only in 2007 did the UTC agree to broaden the scope to provide for emoji, to allow for compatibility with Japanese carrier standards. Since then, emoji have exploded in popularity.

Their encoding, surprisingly, has been a boon for language support. The emoji draw on Unicode mechanisms that are used by various languages, but which had been incompletely implemented on many platforms. Because of the demand for emoji, many implementations have upgraded their Unicode support substantially. That means that implementations now have far better support for the languages that use the more complicated Unicode mechanisms.

Yet emoji remain quite different in kind than other Unicode symbols, which is reflected in fundamental differences in the way that new ones are encoded. It is important for people to understand those differences.

With normal characters, the UTC looks for evidence of usage as text. Proposers need to establish that there is some reasonable body of printed text, either modern or historic, that uses that character. They also need to establish that the usage is not simple glyph variation: we don't encode different characters for variants like {A, A, A, 𝐀, 𝛬, 𝘈, ...}. In theory, this makes for a relatively closed set of possible characters. A detailed character proposal must be supplied, following the directions at Submitting Character Proposals.

Emoji are much different than regular pictographic characters. They are colorful, playful representations of persons, places, or things — and combinations of those (such as a person riding a bicycle). For emoji, rather than look for evidence of existing textual use — since emoji effectively cannot exist in text until they are encoded — we look for evidence of likely high usage once they are encoded, plus a number of other factors. Like normal characters, there is a detailed proposal that must be supplied, following the directions at Submitting Emoji Proposals — however, that proposal is very different from the regular character proposal, reflecting the differences between them.

Emoji are effectively unlimited in variety. In theory, one could have emoji for 339 breeds of dogs, or 10,000 species of birds, and even variants of those (a large female Welsh Harlequin duck, looking over its right shoulder with an egg in the foreground).

Yet that is only in theory: in practice there are many limitations — just not the same limitations as for other Unicode symbols.

1. The only purpose to emoji is if they are widely deployed by major vendors. If not, there is no desire or ability to burden Unicode with pictographic symbols that are not ever "emojified".
2. The major vendors have indicated that they want to hold to an emoji "budget" each year of

about 50-70 new characters. Each additional emoji can be a burden on memory and UI usability — the memory impact is especially important for mobile devices in emerging markets.

3. There is always the option of using emoji-style images (aka stickers) for more specific objects. That is another reason to keep to an emoji budget; every Unicode character is encoded forever, and if emoji go out of style, there is no desire to have an excessive number of them.
4. The process for encoding needs to balance a number of factors. (See Submitting Character Proposals). High among those is prospective usage — if a proposed emoji is not going to be used often by millions of people, then it is taking a slot in the budget that could be occupied by a more popular emoji. Another important feature is breadth: when there are multiple variants of an emoji, the usage just tends to be split among them, while a new kind of emoji permits new kinds of expression.
5. The consortium also tends to roll out small initial sets of new types of characters, such as gender-neutral forms, so that it can assess the frequency of usage before adding more of that type.
6. The consortium has developed a submission process that is is open to anyone, developing factors for encoding that can be applied as objectively as possible to each proposal. Those factors are also applied to internal proposals from Unicode members, and to proposals from liaison members. (That process is open to improvement; the Emoji subcommittee welcomes proposals for improvements).

Non-emoji pictographic characters are normally limited to sets (such as Dingbats) that were encoded for compatibility. New non-emoji pictographic characters are subject to the same process as new letters or other symbols: common use as plain text characters in some body of literature, following Submitting Character Proposals.

Unicode is not open to all possible graphic images as non-emoji pictographs. The Unicode Consortium doesn't approve non-emoji pictographic characters simply to fill in perceived gaps, such as fleshing out a complete taxonomic classification of animal species or varieties.