Unicode-Specified Emoji Customizations

To: UTC
From: Mark Davis, Peter Edberg, Emoji Subcommittee
Date: 2016-01-19
Working Draft: https://goo.gl/5Gbasb

There are many requests for variants of Unicode emoji. This document provides a concrete proposal for a mechanism that could be used for such customization, using the TAG characters. It is based on the discussions in the last UTC meeting (see also L2/16-009). This approach allows us to handle certain kinds of needs without requiring a large number of new encoded characters, and without needed the long lead time for Unicode releases. The mechanism has been designed to be extensible for the future.

The end target for the proposal is for additions to TR51 v3.0, and would include additional emoji properties. While the eventual target for this work would be TR51, for clarity in discussion this material is presented as a single document—not as a list of the precise changes for that document—since it is anticipated that this might first go out as a PRI.

While this mechanism is not required to be synchronized with Unicode 9.0, it would be useful to have it released in around the same timeframe. We would want to allow time for people to assess any implementation issues.

Contents

Semantics

The Unicode emoji customization mechanism is used to request an alternate rendering of a particular emoji character. It is only used where the base emoji character is in some way generic,

and the customization would be considered a variant of that base character. For example, it could be used to indicate that the MAN emoji should be shown with red hair, but not that the MAN emoji be shown as a cup of tea.

Note: one by-product of the direction of this work is that the Emoji SC and UTC should focus primarily on "generic" emoji characters, rather than very specific versions.

Overall Syntax

The customization syntax uses the 95 invisible TAG characters:

U+E0020..U+E007E (TAG SPACE..TAG TILDE).

These correspond to ASCII characters, and may be referenced by an abbreviated name of the form **Tag-<single-ascii-character>**, such as Tag-U for U+E0055 TAG LATIN CAPITAL LETTER U. In examples, where clear, they can also be represented simply by the corresponding ASCII letters. The regex characters ?, *, + have their normal meaning.

In addition, there are the following special terms:

Term	Characters	Description
tag-base	[:emoji=yes:]	any single emoji character (w/o Regional_Indicators)
tag-term	U+E007E TAG TILDE	terminating Tag (see Review Note)
tag-keyChar	Tag-A..Tag-Z	Tag characters corresponding to uppercase letters: [A-Z]
tag-valChar	U+E0020..U+E0040, U+E005B..U+E007D	Tag characters that are neither tag-term nor tag-keyChar.

Review Note: if we un-deprecated U+E007F CANCEL TAG in Unicode v9.0, we could use that for the terminator, which would be a more natural choice.

ED-14a. emoji tag sequence

A sequence consisting of an emoji character followed by one or more non-terminating TAG characters, followed by a terminating TAG character.

```
emoji tag sequence  :=  tag-base tag-setting+ tag-term
tag-base            :=  emoji_character
tag-setting         :=  tag-key tag-value
tag-key             :=  tag-keyChar+
tag-value           :=  tag-valueChar+
```

Further Constraints:

- The entire sequence, including the tag-base and tag-term is limited to 16 characters.
- Within the tag sequence, all keys must be in sorted order, and no key can occur twice.
- Within the tag sequence, no key can occur twice.
- Further structure of the tag-value is determined by the tag-key.

Thus—in ASCII terms—each key is a tag sequence that consist of one or A..Z characters. So “A” is a key, as is “AB” or “ZZ”. The value is any sequence of one or more characters from <space> to “}”. A setting could be, for example “Fusca”. The interpretation of each tag-setting depends on the tag-base, tag-key, and tag-value.

Example: <tag-base>ABxAy<tag-term> is invalid, as is <tag-base>AxAy<tag-term>.

Review Note: the length (currently 16) is up for discussion, but everyone is agreed that we need a fixed limit.

Review Note: the committee considered whether to have the tag-term or not. For font lookup, people didn’t want to have to backup when a non-TAG character is seen. To mark the end, it would be sufficient to forbid sequences that are prefixes of other valid sequences. The easiest way to ensure that is to always have a terminator. In some cases, it may not be necessary, but it makes validity checking much easier.

This defines the well-formed emoji tag sequences. However, the only currently *valid* sequences are those defined in the following sections. All others are reserved for future use. Only certain tag-keys are valid for a given tag-base, and only certain tag-values are valid for the given tag-key. The tag-value may also have internal syntax.

Where the emoji tag sequence is not supported by an implementation, the tag characters are invisible and occupy no space. The emoji character is to be displayed as if there were no tag characters following it.

Review Note: Should we recommend that unsupported tag sequences have a special display, eg dotted rectangle overlay?

With respect to emoji modifiers and ZWJ sequences, an emoji tag sequence behaves as a single emoji character; an emoji modifier that affects the customized emoji should follow the complete sequence representing the customized emoji, and in a sequence such as <Char ZWJ Char TAG+> the TAG sequence applies only to the second character in the sequence.

The TAG sequences also request emoji presentation, so there is no need for a variation selector (nor is one permitted). However, the tag sequence can be followed by combining marks.

Review Notes:

- *Would it be simpler to restrict this so that the TAG characters can only be at the end of an emoji sequence?*

Flags

These customizations allow for additional flag characters, such as the flag of Scotland or California. They are limited to Unicode subdivisions (which generally correspond to ISO 3166-2 subdivisions), and valid 3-digit UN codes.

Syntax

Tag-Base: **U+1F3F3 WAVING WHITE FLAG**
Tag-Key: **Tag-V**
Tag-Value: **(Tag-0..Tag-9, Tag-a..Tag-z)+**

Further Constraints:

1. The Tag-Value must be a specification of either a valid Unicode subdivision_attribute or a valid 3-digit unicode_region_subtag, as per CLDR.
2. They are in canonical form—only lowercase letters and number TAGs: [0-9 a-z].
3. Like the current regional indicators, these can request an image for whatever is currently the flag of the specified subregion. They are *not* intended to provide a mechanism for versioned representations of any particular flag image.

Example: <U+1F3F3>Vgbsct<tag-term> requests the flag for the subdivision "gbsct", which represents Scotland. Note that there is no hyphen, and it is all lowercase, unlike the canonical form for ISO subdivisions ("GB-SCT").

Review Note: if we allowed these flags on the second of a Regional_Indicator pair, the fallback could be the country flag.

Attributes

The attributes are used to request the display of an emoji character to have a particular attribute. There are currently 3 supported types of attributes:

- Hair color (eg, ginger, blonde, brunette, black, gray, white)
- Gender (eg, make a runner be male vs female)
- Direction (for faces with direction, hand gestures, vehicles, pistol, etc.)

Syntax

Tag-Base: **Gender_Base | Hair_Base | Direction_Base**
Tag-Key: **Tag-U**
Tag-Value: **(Gender_Attribute | Hair_Attribute | Direction_Base)+**

The **Tag-Base** and **Tag-Setting** values are listed below. A **Tag-Base** character can be in any of the listed classes. Each class has a set of valid attributes. Additional classes and attributes may be added in future versions.

Further Constraints:

- The **Tag-Value** characters must be in code point order, without duplicates.
- The **Tag-Value** must not contain two characters that are valid for the same class: for example, it cannot contain two **Gender_Attribute** characters.

Additional emoji properties would be added to Annex A: Emoji Properties and Data Files in support of these, with a new data file (eg, emoji-attributes.txt) using the standard format, eg:

```
26F9 ;    Emoji_Gender_Base     # PERSON WITH BALL
1F3C3 ;   Emoji_Gender_Base     # RUNNER
1F3C4 ;   Emoji_Gender_Base     # SURFER
1F3CA ;   Emoji_Gender_Base     # SWIMMER
1F3CB ;   Emoji_Gender_Base     # WEIGHT LIFTER
1F46E ;   Emoji_Gender_Base     # POLICE OFFICER
...
```

There is an attached text file that lists these. If this proposal is accepted, then the list should be presented for public review and refinement before release. Any Unicode 9.0 emoji candidates that are withdrawn in the 2016Q2 UTC meeting would be removed, of course.

Review Note: We could use different tag-keys for the different attributes, eg "G" for gender, "H" for hair, "D" for direction; or "UG", "UH", "UD". That would be conceptually simpler, but longer.

Gender_Base

Those characters that are commonly shown with a particular gender.

Unicode itself doesn't normally specify the gender for emoji characters: it is RUNNER, not MAN RUNNING; POLICE OFFICER not POLICEMAN. The exceptions are where there is corresponding character of the other gender (MAN vs WOMAN), or where the character was encoded for compatibility with the original Japanese emoji or some other source. Those compatibility characters include items like the list below (possibly excluding the snowmen).

	U+2603	SNOWMAN
	U+26C4	SNOWMAN WITHOUT SNOW
	U+1F470	BRIDE WITH VEIL
	U+1F46F	WOMAN WITH BUNNY EARS
	U+1F472	MAN WITH GUA PI MAO
	U+1F473	MAN WITH TURBAN
	U+1F482	GUARDSMAN
	U+1F385	FATHER CHRISTMAS
(9.0 candidates)		
	U+1F57A	MAN DANCING
	U+1F934	PRINCE
	U+1F935	MAN IN TUXEDO
	U+1F936	MOTHER CHRISTMAS

However, for realism vendors typically pick a particular gender to display, even for the neutral characters. The Gender attributes allow vendors to display both gendered forms.

Review Note: If this were adopted in time for Unicode 9.0, we could possibly drop those 9.0 emoji candidates (Prince , Man Dancing , Mother Christmas) that were included only to pair with Princess, Dancer, Father Christmas. Otherwise we'd remove Princess, Dancer, Father Christmas from the above. For the 9.0 sports figures, we could drop those where we have consensus to only

show neutral images. We need to look at whether JUGGLING might be a person, also.

Gender_Attribute

Tag-m	Male appearance
Tag-f	Female appearance
Tag-n	Gender-neutral: neither male nor female appearance

These attributes are to mark *appearance*, and not gender identification. Neutral doesn't mean the default (untagged) presentation, which could be any of these three; it means a specifically a gender-neutral presentation.

Examples:

Male Runner: <U+1F3C3>Um<tag-term>
Female Runner: <U+1F3C3>Uf<tag-term>

Hair_Base

These are characters that are commonly shown with some visible hair.

Review Note: For the 9.0 sports figures, we could drop those where we have consensus to only show neutral images (with no visible hair).

Hair_Attribute

Tag-k	Blac**k**-haired
Tag-s	Blond(e) [also **S**andy-haired]
Tag-b	**B**runet(te)
Tag-g	Redhead [**G**inger]
Tag-w	Gray-haired [also **W**hite-haired]
Tag-d	Bal**d** (no hair)

Review Note: With single letters for hair, it is difficult to be mnemonic, but that doesn't matter to end-users anyway. But other suggestions are welcome.

There are hundreds of possible distinctions among hair color, but because emoji are presented with a "cartoon" style it suffices to have just a few broad choices. This list is taken from the US Online Passport form (BLACK, BLONDE, BROWN, RED, GRAY), with the addition of Bald (no hair). This also matches the UN Grounds Pass application, and is similar to hair color options on other forms, such as driver's licences.

Example:

Red-haired Female Runner: <U+1F3C3>Ugm<tag-term>
Note that the characters "gm" must be in sorted order.

Direction_Base

These are characters typically presented with a direction that may be semantically significant in sequences of emoji, in that they would point in the direction of an action. Unfortunately, for historical reasons the "default" direction is fairly arbitrary; you see some emoji pointing left ▫ and others pointing right ▫.

So, for example, when someone is commenting on a crime movie, for "the detective shot the policeman" it would be more natural to see:

Rather than to see the following, which looks like the detective might be committing suicide:

A (draft) set of Direction_Base characters is listed below.

Note: There are many possible characters that could usefully have direction applied to them. However, we should start out with just small number of base characters, and expand as necessary.

Direction_Attribute

Tag-r	Point-Right
Tag-l	Point-Left

The Tag directions are to have a mirrored effect in a bidi context. All emoji characters are Bidi_Class=Other_Neutral (except for the enclosed alphanumerics).

Example:
Blond Male Runner facing Right: <U+1F3C3>Ugmr<tag-term>
Note that the characters "gmr" must be in sorted order.

Private Use

Private Use tag sequences are for closed interchange within a given system. As with private use codes in general, there is always a danger of collision between different definitions. Any key starting with Tag-X qualifies, and any tag-value.

Syntax

Tag-Base:	**emoji-character (any)**
Tag-Key:	**Tag-X tag-keyChar***
Tag-Value	**tag-value**

Implementations should consider the use of any of the thousands of *private use* Unicode characters instead. However, the advantage of the Private Use emoji customizations is that there is a fallback to the base emoji character, instead of showing a black box.

Review Note: should we use Tag-Z, the last tag?

Review Note: Add more warnings about the use of these sequences from other documents.

Implementation Notes

For subdivision flags, as discussed earlier, there is ***no*** requirement or expectation that all of them — or even any large subset of them — be supported. It is expected that only a relatively small number would be initially supported broadly. (Side note: see the informative TED talk on flag design: The Worst-Designed Thing You've Never Noticed.)

The hair color, gender, and emoji-modifiers could be mixed, resulting in a combinatorial explosion of glyphs in fonts. It is anticipated that only certain combinations would be supported generally. One implementation approach to dealing with the combinatorial explosion is the "**Mx Potato Head**" approach, whereby glyph pieces are assembled for a particular image. For example, there might be some different color hair images that would be appropriate for overlaying on a BOY emoji. These could then be used on each of the BOY images with different skin-tones.

It might be useful to maintain a list of which vendors support which of these tag sequences, something like we already do with the Full Emoji Data chart and other sequence charts.

Review Note: We considered various other models:

- *Additional modifier characters would be possible, but there is a long lead time for defining them, whereas additional settings or values could be added relatively quickly: months not years.*
- *Variation selectors could be used, except that the number is constrained and they are defined as limited to only one per base.*