L2/15-145R

# Proposal for additional regional indicator symbol

*To:* *UTC*
*Date:* *2015 May 7*
*From:* *Unicode emoji subcommittee*
*Link* *http://goo.gl/lbs38F*

---

Unicode 6.0 added 26 REGIONAL INDICATOR SYMBOLS:

```
1F1E6;REGIONAL INDICATOR SYMBOL LETTER A;So;0;L;;;;;N;;;;;
...
1F1FF;REGIONAL INDICATOR SYMBOL LETTER Z;So;0;L;;;;;N;;;;;
```

The code chart annotation for these says:

> **Regional indicator symbols**
> *These characters can be used in pairs to represent regional codes. In some emoji implementations, certain pairs may be recognized and displayed by alternate means; for instance, an implementation might recognize F + R and display this combination with a symbol representing the flag of France.*

And the book text is as follows:

> ***Regional Indicator Symbols.*** The regional indicator symbols in the range U+1F1E6..U+1F1FF can be used in pairs to represent an ISO 3166 region code. This mechanism is not intended to supplant actual ISO 3166 region codes, which simply use Latin letters in the ASCII range; instead the main purpose of such pairs is to provide unambiguous roundtrip mappings to certain characters used in the emoji core sets…The Unicode Standard does not prescribe how the pairs of region indicator symbols should be rendered. In emoji contexts, where text is displayed as it would be on a Japanese mobile phone, a pair may be displayed using the glyph for a flag…

On several systems, these are used to represent from 10 to more than 200 emoji flags corresponding to ISO 3166-1 two-letter codes, which can represent regions such as Isle of Man, Guernsey, and Puerto Rico but not (for example) England, Scotland, Wales, or U.S. States.

On some platforms that support a number of emoji flags, there is substantial demand to support additional flags for the following:

- “Country subdivisions” such as England, Scotland, Wales, U.S. states and Canadian provinces. ISO 3166-2 defines codes for these which consist of a two-letter ISO 3166-1 code followed by hyphen and then a subtag of one to three alphanumeric characters (the possible subtags depend on the ISO 3166-1 code). For example,

"GB-SCT" for Scotland, "GB-WLS" for Wales, "US-DE" for Delaware, and "NO-18" for Nordland (in Norway).

- Certain supra-national regions, such as Europe (European Union flag) or the world (e.g. United Nations flag). These can be represented using UN M49 3-digit codes, for example "150" for Europe or "001" for World.

The proposal has two parts

1. Un-deprecate TAG characters E0020-E007E.

2. Define a character as the "base" for a following sequence of TAG characters that specifies a region or subregion to be represented using a sequence of TAG characters. There are two possibilities for the base character:
    a. Preferred: Use the Unicode 7.0 character WAVING WHITE FLAG: `1F3F3;WAVING WHITE FLAG;So;0;ON;;;;;N;;;;;`
    The advantage is no new characters need be encoded.
    b. Alternate: Encode one more regional indicator symbol immediately preceding the existing symbols: `1F1E5;REGIONAL FLAG BASE;So;0;L;;;;;N;;;;;`
    The chart glyph for this could be a flag in a dotted square:  .

The base plus TAG characters make it possible to designate flags using ISO 3166-2 codes and UN M49 codes as follows:

- Flags corresponding to UN M49 codes are represented using the base character followed by three TAG DIGIT characters for the M49 code.
- Flags corresponding to ISO 3166-2 codes are represented using the base character followed by a sequence of TAG characters corresponding to the ISO 3166-2 code.

Note that this scheme *could* also be used to designate flags corresponding to ISO 3166-1 two-letters codes, using the base character followed by two TAG LATIN CAPITAL LETTERs. This alternative to the use of paired REGIONAL INDICATOR SYMBOL LETTERs to designate ISO 3166-1 codes has better inherent behavior for text break. The committee needs to consider whether such usage should be considered valid, but the remainder of the proposal assumes that it is valid.

Using the following notation —

B designates the chosen base character (U+1F3F3 or new U+1F1E5)
TL designates a TAG LATIN CAPITAL LETTER (A..Z)
TD designates a TAG DIGIT (ZERO..NINE)
TH designates TAG HYPHEN-MINUS

— a well-formed sequence for for designating flags for ISO 3166-1, 3166-2 or UN M49 codes would be

B ((TL{2} (TH (TL|TD){3})?) | (TD{3}))

For stable and non-redundant representation of regions and regional subdivisions using these symbols, some guidelines are useful for determining the validity of such sequences (beyond mere well-formedness):

- When representing ISO 3166-1 or UN M49 regions, only those codes that are valid for the LDML unicode_region_subtag should be used (this prevents multiple representation of many regions which have both an ISO 3166-1 code and a UN M49 code, and provides management of code deprecation, etc.)
- In CLDR 28, LDML will define a unicode_subdivision_subtag which also provides validity criteria for the codes used for regional subdivisions (see CLDR ticket #8423). When representing regional subdivisions using ISO 3166-2 codes, only those codes that are valid for the LDML unicode_subdivision_subtag should be used.

The TAG characters have general category value Cf and line break property value CM. Consequently, either proposed base character followed by a sequence of TAG characters is already treated as a unit for word, sentence, and line break. Grapheme break property values and rules would need some adjustment; until those are updated in UAX 29, implementations could use a tailored grapheme break to handle these correctly.