

Proposal for additional regional indicator symbols

To: UTC
Date: 2015 May 4
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. These can be represented using ISO 3166-2 codes 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).

- 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.

The proposal is to add 46 more regional indicator symbols in the code space immediately before the existing symbols:

```
1F1B8;REGIONAL SUBDIVISION SYMBOL DIGIT ZERO;So;0;L;;;;N;;;;;
...
1F1C1;REGIONAL SUBDIVISION SYMBOL DIGIT NINE;So;0;L;;;;N;;;;;
1F1C2;REGIONAL SUBDIVISION SYMBOL LETTER A;So;0;L;;;;N;;;;;
...
1F1DB;REGIONAL SUBDIVISION SYMBOL LETTER Z;So;0;L;;;;N;;;;;
1F1DC;REGIONAL INDICATOR SYMBOL DIGIT ZERO;So;0;L;;;;N;;;;;
...
1F1E5;REGIONAL INDICATOR SYMBOL DIGIT NINE;So;0;L;;;;N;;;;;
```

For REGIONAL INDICATOR SYMBOL DIGITs the chart glyph would be a digit in a dotted square; for REGIONAL SUBDIVISION SYMBOLs it could be the corresponding symbol within a double dotted square, to indicate a contained subregion—for example: . The properties for the new characters are the same as for the existing REGIONAL INDICATOR SYMBOLs.

The regional indicator symbol digits and the subdivision symbols make it possible to represent ISO 3166-2 codes and UN M49 codes as follows (the initial idea was from Deborah Goldsmith):

- UN M49 codes are represented using three REGIONAL INDICATOR SYMBOL DIGITs.
- ISO 3166-2 codes are represented by two REGIONAL INDICATOR SYMBOL LETTERs, followed by a one-to-three-character subtag consisting of REGIONAL SUBDIVISION SYMBOLs (letters and/or digits per the ISO 3166-2 codes).

REGIONAL SUBDIVISION SYMBOLs need to be encoded separately from the existing REGIONAL INDICATOR SYMBOLs so that

- existing processes do not interpret pairs of letters in subtags as representing ISO 3166-1 codes, and
- there is a clear boundary between letters and digits that are part of a subtag and letters or digits that are part of any immediately following REGIONAL INDICATOR SYMBOL sequence (two letters or three digits).

Then using the following notation —

RL designates a REGIONAL INDICATOR SYMBOL LETTER;

RD designates a REGIONAL INDICATOR SYMBOL DIGIT;

RS designates a REGIONAL SUBDIVISION SYMBOL

— a well-formed regional indicator sequence will have the following syntax:

(RL{2} | RD{3}) (RS{1,3})?

For stable and non-redundant representation of regions and regional subdivisions using these symbols, some guidelines are useful:

- When representing regions using REGIONAL INDICATOR SYMBOLS, 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 subdivision subtags using REGIONAL SUBDIVISION SYMBOLS, only those codes that are valid for the LDML `unicode_subdivision_subtag` should be used.

There is currently ambiguity in determining the boundary of a sequence of REGIONAL INDICATOR SYMBOLS. Thus we recommend that any sequence of two REGIONAL INDICATOR SYMBOLS be terminated with ZERO WIDTH NON-JOINER. The new SUBDIVISION characters do not increase the ambiguity; in fact, ZWJ is not needed after any SUBDIVISION character.

With the new characters, some UAX #29 changes would be needed, at least as follows:

http://unicode.org/reports/tr29/#Grapheme_Cluster_Boundaries

any sequence of Regional_Indicator (RI) characters
=> certain sequences starting with Regional_Indicator (RI) characters

http://unicode.org/reports/tr29/#Regex_Definitions

RI-Sequence	:= Regional_Indicator+ Subdivision_Indicator*
-------------	---

http://unicode.org/reports/tr29/#Grapheme_Cluster_Break_Property_Values
http://unicode.org/reports/tr29/#Table_Word_Break_Property_Values

Regional_Indicator: add the new digits

Subdivision_Indicator: add a new property value, with the new digits/letters

<http://unicode.org/reports/tr29/#GB8a>
<http://unicode.org/reports/tr29/#WB13c>

Do not break between regional or subdivision indicator symbols

GB8a. Regional_Indicator × (Regional_Indicator | Subdivision_Indicator)

Subdivision_Indicator × Subdivision_Indicator

(proposal summary form below to be completed pending discussion in UTC)

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646

Please read [Principles and Procedures Document \(P & P\)](#) for guidelines and details before filling this form.

A. Administrative

1. Title: Additional regional indicator symbols
2. Requester's name: Unicode emoji subcommittee
3. Requester type (Member body/Liaison/Individual contribution): ???
4. Submission date: 2015-05-01
5. Requester's reference (if applicable):
6. Choose one of the following:

This is a complete proposal: Yes

...

B. Technical – General

1. Choose one of the following:

...

b. The proposal is for addition of character(s) to an existing block: Yes

Name of the existing block: Enclosed Alphanumeric Supplement

2. Number of characters in proposal: 46

...