L2/15-210

PRI 299 feedback and mailing list discussion

To: *UTC*
Date: *July 28, 2015*
From: *Peter Edberg*

PRI 299 and related unicode list discussion, organized by primary topic in a thread-view style. Topics:

Subtag validity

Date/Time: Tue Jun 30 14:18:36 CDT 2015
Name: Doug Ewell
Report Type: Public Review Issue
Opt Subject: PRI #299
I support this proposal. I have the following questions:

1. The existing RIS-based flag mechanism is based on ISO 3166-1 (TUS 7.0 §22.10). In this proposal, "valid" tag sequences would instead be determined by CLDR data and LDML specification. Is there any precedent for CLDR to define the validity of Unicode character sequences?

> **On Jul 2, 2015, at 9:10 AM, Mark Davis wrote:**
> We already have, in tr51, the unicode_region_codes being used for validity testing of flags:
> http://unicode.org/reports/tr51/#Encoding
> http://unicode.org/reports/tr51/#Flags
>
> > **On Jul 2, 2015, at 10:33 AM, Doug Ewell <doug@ewellic.org> wrote:**
> > the second of which (Annex B) says:
> > "The valid region sequences are specified by Unicode region subtags as defined in [CLDR], excluding those that are designated private-use or deprecated in [CLDR]."
> >
> > In that case, the wording in TUS needs to be corrected, because TUS 7.0 §22.10 says:
> > "The regional indicator symbols in the range U+1F1E6..U+1F1FF can be

> > > used in pairs to represent an ISO 3166 region code."
> > >
> > > It doesn't say anything about valid pairs being defined by CLDR instead of ISO. I wonder how many users actually know this.
>
> Those are typically the same as the ISO codes, but do add XK
> http://unicode.org/reports/tr35/#unicode_region_subtag
>
> > **On Jul 2, 2015, at 10:33 AM, Doug Ewell <doug@ewellic.org> wrote:**
> > So QO, QU, and ZZ would be excluded, since those are private-use in BCP 47 and hence also in CLDR. But XK is included, even though it is also private-use. Is this correct? Can an application tell that XK is in and the others are out, just by looking at CLDR data?
> >
> > Also, I assume all of the same include/exclude rules apply both to RIS combinations and to PRI #299-style flag tags. Please let me know if that's not true.
>
> Peter E: As for precedent, Technically this is proposing to have language in UTR #51 defining the validity of sequences based on criteria provided by the CLDR LDML spec and CLDR data. It is really just a way of more precisely specifying the use of ISO 3166-1 and 3166-2 codes.

2. What is the policy on generating flag tags with deprecated unicode_region_subtag or unicode_subdivision_subtag values, such as "[flag]UK"? How "discouraged" would such a tag be? Should tools allow users to create such a tag?

> **On Jul 2, 2015, at 9:10 AM, Mark Davis wrote:**
> CLDR treats UK as deprecated. When a code is deprecated, we strongly discourage its use in new data, but normally allow it for old data. But the UK is somewhat different, since it really shouldn't ever be valid as it stands. The purpose for UK in CLDR metadata is so that locale ID canonicalization can map en-UK (which occurs quite often) to en GB, and so on. (We do this also for overlong codes like eng-GB => en-GB.)
>
> But you're right; we need to be able to distinguish this case (and ones like it.) I filed
> http://unicode.org/cldr/trac/ticket/8736
>
> > **On Jul 2, 2015, at 10:33 AM, Doug Ewell <doug@ewellic.org> wrote:**
> > OK, so UK is not valid in RIS combinations or flag tags either. Glad to see that clarified.

3. The subdivisions.xml file contains a "subtype" hierarchy, reflecting the "parent subdivision" relationship in ISO 3166-2. So region 'FR' contains subdivision 'J' (Île-de-France), which itself contains subdivision '75' (Paris). Is there any significance to the "subtype" hierarchy as far as flag tags are concerned, or are "[flag]FRJ" and "[flag]FR75" equally valid?

> **On Jul 2, 2015, at 9:10 AM, Mark Davis wrote:**
> No, there isn't. But see also E.5 in
> http://www.unicode.org/review/pri299/pri299-additional-flags-background.html
>
> > **On Jul 2, 2015, at 10:33 AM, Doug Ewell <doug@ewellic.org> wrote:**
> > Right, clearly flags don't exist for many of the subdivisions. But I'm not sure this is the same question as whether the three-level hierarchy is relevant. In my example, Île-de-France and Paris both have flags, and they aren't the same. (Wikipedia says the Île-de-France flag is "non-official and unused," but they do have a page for it, and in any case there are probably better examples.)

4. The entry for "001" in subdivisions.xml contains each of the two-letter codes for regions (countries) that have their own subdivisions. This is less than the set of all regions; for example, Anguilla (AI) does not have ISO 3166-2 subdivisions and so is not listed. This implies that a tag like "[flag]001US" is valid (and equivalent to "US" spelled with RIS, which is preferred) but "[flag]001AI" is not valid. Is this intended? If not, can it be clarified?

> **On Jul 2, 2015, at 9:10 AM, Mark Davis:**
> Good catch, the 001 shouldn't even exist in the subdivisionContainment. This is now fixed in trunk.
> (The subdivision addition will only be final in September, so feedback on it now would be great. People can file tickets at http://unicode.org/cldr/trac/newticket)

5. Will any preliminary examples of CLDR 4-character subdivision codes be made available before any such codes are actually assigned?

> **On Jul 2, 2015, at 9:10 AM, Mark Davis wrote:**
> The only purpose for the 4-character subdivision codes is stability. So let's suppose that Colorado decides to join Canada (thereby deprecating CO in ISO 3166-2), and British Columbia decides to join the US (getting the code CO in ISO 3166-2). In that case, CLDR would keep the old code CO (but deprecated) and create a new 4-letter code for BC, such as XXCO. This is just for illustration, of course, I've heard no rumors about either political shift...
>
> > **On Jul 2, 2015, at 10:33 AM, Doug Ewell <doug@ewellic.org> wrote:**
> > Thanks for the 'XXCO' example; this is different from tending toward 'COXX' and was what I was looking for.
> >
> > The exact scenario would not apply, of course, due to the agreement to keep subdivision codes unique across the US/Canada border. I'd suppose this would be preserved, and 3166-2 would assign US-BC to "British Columbia as US state," and there would be no coding conflict to resolve. But again, additional examples could easily be dreamed up: replace BC with the Central Abaco region of the Bahamas (currently BS-CO), which isn't that far away.

Date/Time: Thu Jul 2 10:05:48 CDT 2015
Name: Doug Ewell
Report Type: Public Review Issue
Opt Subject: PRI #299

6. What is the policy on generating flag tags with unicode_region_subtag values corresponding to private-use BCP 47 subtags, other than those given special semantics by CLDR? Are they invalid or merely discouraged? Should tools allow users to create such a tag? Is there any provision for a "private agreement," similar to that defined in Unicode for PUA usage?

> **On Jul 2, 2015, at 9:10 AM, Mark Davis wrote:**
> We'll have to address that. My view is that they should not be valid: if someone wants a PU flag, of any source, they have over 130,000 Unicode PU character s to play with.
>
> > **On Jul 2, 2015, at 10:33 AM, Doug Ewell <doug@ewellic.org> wrote:**
> > I concur, and this is consistent with Annex B.

Base character

Date/Time: Thu Jul 2 16:17:17 CDT 2015
Name: Leo Broukhis
Report Type: Public Review Issue
Opt Subject: PRI #299

Using default-ignorable characters results in an undesirable fallback behavior: a U+1F3F3 WAVING WHITE FLAG by itself is not informative.

To quote, "This base character is a visible spacing character that suggests a flag, so that implementations that do not support the TAG characters have an indication that a flag is present."

However, whenever a flag is used, the primary intention is to indicate the entity or concept represented by a flag, therefore the preferred fallback behavior would be to lose the flag form while keeping the identifying information visible. This is better achieved with a scheme using visible characters, like already existing regional indicator symbols.

> **On Jul 2, 2015, at 1:58 PM, Leo Broukhis <leob@mailcom.com> wrote:**
> What I don't like about PRI #399 is its proposing to use default-ignorable characters. On a non-vexillology-aware platform, I'd like to see something informative, albeit not resembling a flag, but indicative of the intention to display a flag, like RIS can be, as opposed to nondescript white flags.
>
> **On Jul 3, 2015, at 4:28 PM, Asmus Freytag (t) <asmus-inc@ix.netcom.com> wrote:**
> My concern is that there are good reasons to have more than just a rectangular static "flag" plan image. There are triangular flags as well, and a flag that's drawn to be flying gives a different "festive" or "dynamic" image than a static one.
>
> In my view, the choice of a single base character is what makes the proposal unnecessarily limited. A new "WHITE FLAG" should be added, that is not "WAVING" to serve as the canonical base character, leaving any other flag shapes as base characters for flags of that shape.
>
> **On Jul 6, 2015, at 8:18 AM, Doug Ewell <doug@ewellic.org> wrote:**
> I think a useful bit of feedback on PRI #299 would be to inquire whether it is, in fact, a design goal to handle this use case of transparency of the individual letters on platforms, rendering engines, and/or fonts that don't support flag-tag composition. (Please, not "non-vexillology-aware." None of these platforms studies or analyzes flags. They assemble multiple characters into a single image.)
>
> If transparency on flag-tag-unaware platforms is not a design goal, it might be difficult to make the case that default-ignorable tag characters are a poor choice because they don't support transparency.
>
> **On Jul 6, 2015, at 1:40 PM, Doug Ewell <doug@ewellic.org> wrote:**
> Leonardo Boiko <leoboiko at namakajiri dot net> wrote:

I think a waving white flag is an emoji symbol for "truce/surrender/come in peace", whereas a white rectangle doesn't easily transmit the same idea.

I don't know how many other flags have different semantics depending on whether they are waving or not. I note that neither RIS pairs nor PRI #299 sequences can encode a plain white flag (but of course the user can simply choose between U+2690 and U+1F3F3 for that).

I hear Asmus's concern about using WAVING WHITE FLAG as the base character for emoji flags which might not be depicted as waving. However, in that case the solution would be to choose a different, *single* base character.

Date/Time: Mon Jul 20 17:19:11 CDT 2015
Name: Doug Ewell
Report Type: Public Review Issue
Opt Subject: PRI #299: base character

In response to objections raised on the public mailing list, I suggest that the exact appearance of the base character (white vs. black flag, still vs. waving, rectangular vs. triangular, etc.) will not make a significant difference and should not, by itself, cause this tagging mechanism to be delayed until a new character can be encoded, in Unicode 9.0 at the earliest.

In particular, distinct base characters should not be used to distinguish a still flag from a waving flag, because some fonts already show the existing RIS-based flags as waving while others already show them as still, and because there is little or no semantic difference for country and subdivision flags such as those under discussion here.

Extension/Registry mechanism

Date/Time: Tue Jun 30 14:40:22 CDT 2015
Name: Doug Ewell
Report Type: Public Review Issue
Opt Subject: PRI #299

The PRI #299 mechanism is clearly and intentionally oriented toward representing flags of well-defined geopolitical entities.

Any proposal to extend the mechanism to cover the many other types of flags -- for historical regions, NGOs, maritime, sports, or social or political causes -- must be systematic and well-planned, not ad-hoc or haphazard, to assure interoperability and extensibility.

> **On Jul 2, 2015, at 9:10 AM, Mark Davis wrote:**
> Firmly agreed.

The documentation for the PRI #299 mechanism should state clearly that (e.g.) the Confederate battle flag, the Olympic flag, the Esperanto flag, the LGBT rainbow flag, and the naval flags used to spell out "ENGLAND EXPECTS" can be represented only via a proper extension to the mechanism, not by ad-hoc means such as the use of unassigned or private-use combinations. This is at least as important as ensuring the stable coding of geopolitical flags.

> **On Jul 2, 2015, at 9:10 AM, Mark Davis wrote:**
> Yes, again a good point.

Date/Time: Thu Jul 2 17:24:45 CDT 2015
Name: Ken Whistler
Report Type: Public Review Issue
Opt Subject: PRI #299: a registry mechanism

I suggest that the correct way to incorporate a generalized extension mechanism for the PRI #299 scheme of use of tag characters together with CLDR-defined unicode_region_subtags and unicode_subdivision_subtags is as follows:

Define the private use region subtag "XF" as meaning flag pictograph defined in the Unicode Flag Pictograph Registry. (UFPR)

Use a 4 digit number, starting from 0001 and extending (in principle) to 9999 as a unique identifier for registered entries in the UFPR. (Should 10,000 entries prove insufficient, this scheme could later be extended almost indefinitely as A000..A999, B000..B999, etc.)

Use the same syntax as for the regular region/subdivision tag identification of particular flags. So, e.g., for UFPR registration number 0001, a corresponding full tag sequence would be:

<Base, TAG-X, TAG-F, TAG-0, TAG-0, TAG-0, TAG-1>

This would be syntactically completely consistent with the rest of the proposed mechanism. Semantically, it would require only knowing that "XF" branches out to the UFPR for validation, instead of to the CLDR tables of valid region/(contained)subdivision pairs. In all other respects, including fallback to display of the sequence simply by the glyph for the base flag symbol, this mechanism would behave identically.

To support the UFPR registry, initiate a new UTS, structured along the lines of UTS #37, which currently defines the IVD and its maintenance procedures. The new UTS would define the applicable procedures for submission and review of a flag pictograph for registration. It should include fairly strict criteria that would guarantee non-overlap with flag pictographs already representable by the RIS pairs or proposed tag extension mechanism with region/subdivision tags, and to prevent against nuisance, trivial, or duplicate

registrations. It should also have strict requirements for clear submission of a specific representative color glyph (with unencumbered IPR) and clear identification and intended use. Mass submissions by vague reference to external sources would not be allowed, although presumably the registrar could work out streamlined procedures for bona fide submissions of well-defined sets of related flag pictographs -- e.g. a set of maritime flags, signal flags, or such.

Other details would need to be worked out, of course, but I envision a fairly open set of criteria, whereby the registrar would not block registrations based on political, ideological, religious, or other potential biases, except insofar as a proposed registration might clearly run into national or international legal issues (think Nazi flag), or where unencumbered IPR was not demonstrable (think flags for corporations).

To be useful for interchange purposes, the actual UFPR registry would have to be easily accessible online, with its content consisting at a minimum of distinct records showing the registration number, a clear and distinct representative color pictograph, and unambiguous identifying metadata for the registration. It would *not*, however, be a glyph *service*. Implementers and vendors would need to choose which entries they would support in practice, and would be responsible for designing and making available the actual flag pictographs in fonts they would use, consistent with any such pictographs they implement for the other standard (non-registry) mechanisms.

Unicode list - hyphen in syntax

From: Philippe Verdy <verdy_p@wanadoo.fr>
Date: July 1, 2015 at 1:57:49 AM PDT

I oppose this proposal for the simple reason that it thinks hyphen separations are not necessary. Possibly true today but there will be extensions in some future needing more than 2 letters or 3 digits in the primary subtag. even for iso 3166-2 the regional subtags are very likely to change and without separators the extension,s will become ambiguous

---

Unicode list - stability of representation

(Several people mentioned that encoding flags by region is not stable and unambiguous, e.g. “geopolitical entities and flags (as a specific instances of a design, in the heraldic sense) are disjoint. And that using geopolitical codes to refer to these designs is inherently unstable.”)

On Jul 2, 2015, at 10:04 AM, Ken Whistler <kenwhistler@att.net> wrote:

> Re: “In really there's still no standard way to encode flags unambiguously and in a stable way. We'd like to have FOTW (Flags of the World) contributors to propose their own scheme. But it will not be compatible with the current RIS solution or the proposed extension. If ever such standard emerges, it will require encoding a new set of characters.”

The UTC is neither responsible for nor interested in a "standard way to encode flags unambiguously". I suspect one of the reasons this discussion is tending to derail into political topics and too much detail about

particular flags and their stability and the stability of geopolitical entities they represent and yadda yadda, is that people seem ineluctably drawn to the misapprehension that this is all about standard encoding of flags.

It is not.

Rather, it is about a standard way to represent recognizable and interchangeable emoji (colorful little pictographs) of flags, using defined sequences of Unicode characters.

The existing mechanism using regional indicator symbol (RIS) pairs was originally aimed at solving the following problems:

1. Enabling the reliable interchange of the legacy 10 flag emoji from Japanese carrier sets.

2. Enabling the completion of the encoding of emoji to cover the rest of the Japanese carrier sets without all progress dragging to a complete halt as national bodies in SC2 would argue interminably over a "standard way to encode flags unambiguously" in an ISO standard.

3. Dealing with the inevitable hue and cry: "China and Japan and the US got their flag! Why can't I get my country's flag??!"

And it appears that the RIS mechanism succeeded spectacularly well in addressing all of those design goals.

In the middle of last year, for example, there was a major media and internet campaign to "encode the flag of India". Well, the RIS mechanism handled the real issue there just fine -- when the new phones started coming out with support for display and interchange of emoji for flags using the RIS sequences, there was the emoji for the flag of India for everybody to use. Problem solved.

And the problem which was solved was not the determination that the <1F1EE, 1F1F3> RIS sequence "IN" meant precisely the current national flag of India, the saffron, white and green tricolor with the Ashoka Chakra, and *not* any other flag of India (the flag of the Indian army, the flag of the Mughal Empire, the flag of British India, etc.). The RIS sequence "IN" was just mapped to the colorful little emoji glyph for the Indian flag that everybody wanted to interchange.

The Unicode Standard is not a vexillology standard -- nor will it ever be. It is a standard for the encoding and interchange of characters.'

> **On Jul 2, 2015, at 12:09 PM, Doug Ewell <doug@ewellic.org> wrote:**
> Even though I continue to believe there *should* be a vexillology standard for encoding flags as unambiguously as practicable, I'm in strong agreement that this is not a Unicode problem, or a character problem, or even a CLDR problem.
>
> If there were such a standard today, it might make sense for Unicode and/or CLDR to adapt it for the emoji purposes we are discussing here. But there isn't.
>
> **On Jul 2, 2015, at 1:58 PM, Doug Ewell <doug@ewellic.org> wrote:**

> ….A proper coding standard for flags (NOT in scope for Unicode) might have this sort of versioning feature, but even then, I would think the default (unversioned) behavior should be to select the "current" flag, whatever that is.

The *character* problem we are faced with here is that people want to use and interchange colorful little emoji pictographs of various flags in text streams. The RIS mechanism addresses a significant part of that problem, but is not extensible to cover the full scope of the demand.

And what is the scope of the additional demand?

1. The first part can be summed up as: **the flag of Scotland problem.**

In other words, there are a number of high visibility, high demand, widely recognized regional flags that would be interchanged as just more emoji pictographs, if a mechanism for that were available.

People who want to use an emoji for the flag of Scotland just as easily as someone can use an emoji for the flag of Great Britain are not going to accept an argument that says, "Well, we can't do that on your phones because there is no 3166-1 country code registered, so we can't map a Scotland flag emoji glyph to a RIS pair."

Hence the PRI #299 proposal: for an extension mechanism that would address the flag of Scotland problem in a generic and reasonably stable way.

2. The second part can be summed up as: **the rainbow flag problem.**

In other words, there are a number of high visibility, high demand, widely recognized non-governmental flags that would be interchanged as just more emoji pictographs, if a mechanism for that were available.

From the public's point of view, this is another no brainer: if the flag of Japan and the flag of Scotland, why not the rainbow flag??! They aren't interested in the limitations of the underlying representation mechanisms, nor should they be, IMO.

The problem the UTC faces here is that there are a number of reasonable and popular candidates, which the rainbow flag amply exemplifies, for more colorful little emoji pictographs for flags that people would like to interchange -- but there is no obvious and extensible way to do so reliably in terms of sequences of Unicode characters in a plain text stream. The PRI #299 proposal does not extend into this realm, for many of the reasons pointed out by Doug Ewell.

There are a number of potential approaches to address the rainbow flag problem. For example:

a. use private-use characters
b. pursue one-by-one encoding of each newly desired flag pictograph as a symbol
c. extend the unicode_region_subtag and unicode_subdivision_subtag scheme in CLDR to add some new subtag addressing a separate, non-geopolitical hierarchy
d. create a separate extension using TAG characters but with a syntax not dependent on CLDR subtag definitions
e. create a registry of flag entities suitable for representation as emoji, together with a "c" or "d" style syntax
f. something else?

g. do nothing (and perhaps hope that stickers will solve the problem)

If we are to make any progress here in addressing the actual scope of "the rainbow flag problem", I suggest we focus on the details and pros and cons of suggestions like those of a through g above, rather than pursuing more discussion recapitulating the history of the borders of Tibet -- which truly are out of scope here.

On 7/2/2015 5:56 PM, Peter Constable wrote:
For the proposal to be workable requires some means of ensuring stability of encoded representations. The way this would be done would be for CLDR to provide data with all valid sequences --- effectively becoming a registry.

> **On Jul 3, 2015, at 11:23 AM, Ken Whistler <kenwhistler@att.net> wrote:**
> **I** think that is wrong on a couple of grounds.
>
> First, detailed stability of reference to actual defined geopolitical entities or particular detailed flag designs is not *required* for proposal to represent *pictographs* of flags by some sequence of Unicode characters to be "workable". Sure, more stability of reference is desirable. But the current RIS pair mechanism for representing flag pictographs for countries is already "workable" -- it works and is widely deployed and widely used -- without having guarantees that some particular country may not decide tomorrow to change its official flag and hence result in some particular pictographic display being obsolete in some sense, for example.
>
> Second, the horse is already out of the barn regarding the particular data that CLDR would be referring to. This works by reference to the ISO 3166-2 scheme of subdivisions:
>
> https://en.wikipedia.org/wiki/ISO_3166-2
>
> and *that* becomes the registry required for stability of representations, plus whatever grandfathering stability-of-code mechanism BCP 47 adds on top of that. We don't require a further detailed level of registration, I think, to make this workable…