Integration Panel

Disposition of the MSR1 comments by Neha Gupta et al.

Last updated: May 7, 2014

Summary

This document describes a proposed disposition of the comments by Neha GUPTA (et al.) and Akshat JOSHI posted on the ICANN web site as http://forum.icann.org/lists/comments-msr-03mar14/msg00003.html and the addendum at http://forum.icann.org/lists/comments-msr-03mar14/msg00004.html, as part of the Public Comment of the Maximal Starting Repertoire (MSR1) of the Root Zone Label Generation Rule Procedure (http://www.icann.org/en/news/public-comment/msr-03mar14-en.htm).

The dispositions address only comments that request an action by the Integration Panel or a change to MSR1.

1 General Comments

1.1 Comment 1:

Since the issue was raised during the presentation of MSR in Singapore meeting (49th ICANN Public Meeting - IDN Root Zone LGR Public Workshop - Integration & Generation Panels), and since Ethnologue data was used to determine which language is in active use; it is suggested that the concept of EGIDS as used by Ethnologue (https://www.ethnologue.com/about/language-status) could be incorporated within the document as a short note. This would ensure clarity.

1.1.1. Requested Change

Document EGIDS in the Rationale and Overview document.

1.1.2. Disposition

The Integration Panel disposition is to accept this request.

1.2 Comment 3 – on section 5.6

The comment states that the arguments that consider the existing Candrabindu characters to be unstable in light of pending additions is misplaced because "*The scripts Kannada, Malayalam were not using Devanagari Sign Candrabindu in their text prior to this introduction. As per our knowledge, it was nowhere flagged as requirement either from Unicode or from the language community as well.*

Also, in case of Telugu, the script already had a Candrabindu character encoded as (U+0C01 - TELUGU SIGN CANDRABINDU, also known as arasunna). The newly introduced Candrabindu in Telugu is actually a TELUGU SIGN COMBINING CANDRABINDU which is not same as the traditional Candrabindu." and that the re-use of "*Devanagari Candrabindu with these scripts would not have been possible since the*

resulting display would show that the combination is illegal and further would not be acceptable to a native user of the script/language:

Below is an example for reference:

కँ- TELUGU LETTER KA + DEVANAGARI SIGN CANDRABINDU“

1.2.1. Proposed Change

Correct the analysis

1.2.2. Disposition

The Integration Panel disposition is to accept this request.

2 Specific Comments

2.1 Page 7:

Not all scripts of India follow comparable rules of consonant vowel. This is applicable only to scripts based on Brahmi.

Although the section on Confusability is clear an example would help

2.1.1. Requested Change

Correct statement on page 7 and provide examples on Confusability.

2.1.2. Disposition

The Integration Panel disposition is to accept this request and update the document.

2.2 Page 8 et seq. — Use of the word Bengali to indicate the script.

*It has been noted that **Bengali** is used to indicate the script. However, as per the recent update on the Unicode code charts [index] page, the script is depicted as “Bengali and Assamese” Reproducing below the snapshot for ready reference.*

[The comment provides a snapshot from the index page]

Thus, we feel, the use “Bengali and Assamese” instead of Bengali would be more appropriate since this apparently is the new nomenclature adopted by Unicode 6.3. It is requested that the Integration Panel may need to address this point in appropriate text.

2.2.1. Requested Change

Change Bengali to “Bengali and Assamese”

2.2.2. Disposition

Contrary to what the comment implies, the formal name of the script in Unicode continues to be “Bengali” as documented in the Unicode script property [UAX#24]. This is notwithstanding the fact that the index page also provides “Assamese” as an alias to aid users in locating the correct script.

The Integration Panel disposition is to reject the request, but instead add an annotation.

2.3 Page 16 and 20

*It is unfortunate that Sindhi ampersand and postposition have been treated as Punctuations whereas in fact these are the characters that coincide with a given word: '*and*' and postposition: '*in*'.*

2.3.1. Requested Change

No change to the MSR repertoire or WLE Rules; the intended request is unclear.

2.3.2. Disposition

These are formally classified as "symbols" in Unicode (General_Category value). They are so referenced in the MSR Overview and Rationale document.
The disposition of the Integration Panel is to take no action.

2.4 Page 17 — Candrabindu characters

In the case of Devanagari, Candrabindu is required extensively and hence should not be excluded from MSR. [A table of usage for principle languages using the Devanagari script is provided in the comments]

In the case of Bengali and Assamese, Candrabindu is similarly used actively and hence should not be excluded from MSR.

Gurmukhi sign Adak Bindi (U+0A01) is not a candrabindu per se. It is a combination of two nasal markers Addak (U+0A71) and Bindi (U+0A02). It is used very sparingly and may be continued to be removed from the MSR.

In the case of Gujarati Candrabindu is very sparingly used mainly to show loan words e.g. મુઁહ: face

The usage of Candrabindu in this script is slowly spreading though. It may be included in the MSR, an appropriate decision may be taken by the Generation Panel.

In the case of Oriya, as in the case of Devanagari, Candrabindu is actively used and should not be excluded from the MSR. The word for mother is in fact the first word figuring in Oriya primers and hence should not be deprecated.

2.4.1. Proposed Change

Add

- U+0901 DEVANAGARI CANDRABINDU SIGN
- U+0981 BENGALI SIGN CANDRABINDU
- U+0B01 ORIYA SIGN CANDRABINDU

to MSR1, and consider adding U+0A81 GUJARATI SIGN CANDRABINDU

2.4.2. Disposition

The Integration Panel accepts the input presented in the comment as sufficient to establish that the future addition of candrabindu characters in several Indic scripts does not destabilize the use of existing candrabindus. The Integration Panel agrees that there is sufficient justification presented to include the four code points listed above in the MSR. Particularly in the case of the Gujarati candrabindu, the

Integration Panel would look towards a well-argued rationale for why these code points are truly essential for top level domain names (and not merely in order to ensure completeness in covering the orthography).

The Integration Panel disposition is to accept this addition.

2.5 Page 20: OM

*OM should be grouped under **5.9 Code points used for Religious Purposes***
To the list should be added:
U+0A74 : ੴ Gurmukhi Ek Onkar

2.5.1. Proposed Change

Add U+0A74 GURMUKHI EK ONDAR to list in section 5.9 of the MSR Overview and Rationale document.

2.5.2. Disposition

The Integration Panel disposition is to accept this update.

2.6 Page 21 — Resurgent Scripts

It is suggested that a separate section to be provided for resurgent scripts i.e. Scripts which have been revived by a Linguistic Community e.g. Ol Ciki, Meetei Mayek. Hebrew is a classic example of such resurgence and acceptance both in writing and speaking.

2.6.1. Proposed Change

No change to the repertoire, but recognize Ol Chiki and Metei Mayek as resurgent.

2.6.2. Disposition

The term used by Unicode for scripts for which efforts are under way to establish or re-establish use is “aspirational”. The tables in the Overview and Rationale document are based on tables published by Unicode in [UAX#31], where these scripts are listed under “limited use”. For the purpose of the LGR both categories are treated the same. Both are eligible for reevaluation for the MSR, if there is a substantial development that did result in widespread use of the script. So, rather than departing from Unicode’s classification, an annotation on page 11 might be considered that these two scripts are considered resurgent.

The Integration Panel disposition is to accept this request by adding an annotation.

3 Addendum on 02BC

3.1 Comment

Three Indian languages use MODIFIER LETTER APOSTROPHE as part of their character set. These are Boro, Maithili and Dogri.

In all three languages the character resembles an apostrophe mark which is placed flush along with the shirorekha i.e. the top line and cannot be confused with an apostrophe.

When Unicode was approached to encode this character and assign it a code-point with the code block of Devanagari, Unicode decided to allot for this modifier letter the code-point 02BC.
02BC is part of the Spacing Modifier letters range
02B9..02C1; Common # Lm [9] MODIFIER LETTER PRIME..MODIFIER LETTER REVERSED GLOTTAL STOP

Unicode chapter 9… clearly indicates that 02BC is used in three languages using Devanagari script and thus warrants inclusion within the MSR of characters which need to be included.

[The comment include a reproduced section of the Standard describing usage]

It is requested that since these three languages are official languages of India and their joint population is as under as per Ethnologue; 02BC be included in the repertoire.

- *Maithili: 30,000,000 in India (2000 SIL).*
  Population total all countries: 33,890,000. 12,000,000 monolinguals (1998).
- *Dogri: 2,280,000 (2001 census).*
- *Bodo/Boro: 1,330,000 in India (2001 census). Population total all countries: 1,334,380.*

We are fully aware that the introducing U+02BC in the MSR may pose some of the challenges as it is perceived as punctuation, which is not really the case insofar as these three languages are concerned. It is requested that the Integration Panel permit this as a part of the MSR and let the further decision on its inclusion/exclusion be taken by the community in the Generation Panel.

3.2 Request

Add U+ 02BC MODIFIER LETTER APOSTROPHE to the MSR

3.3 Disposition

In the MSR-1 released for public comment on March 3rd, 2014, the Integration Panel documented in the MSR Rationale and Overview document, section 5.7.4, that homoglyphs of non PVALID characters, such as punctuations were also excluded from MSR. By being part of the range U+02B9..U+02C1, the code point U+02BC was excluded from the MSR on the ground that it is a homoglyph of U+0027 (Apostrophe) or its typographical equivalent (U+2019), which are DISALLOWED under IDNA2008. It is worth noting that while U+02BC and U+0027 (or U+2019) are distinct code points, in standard typographical use the representation of an apostrophe and letter apostrophe are indistinguishable; they are homoglyphs.

The Integration Panel considers characters that are homoglyphs of characters assigned to DISALLOWED code points normally unacceptable in the root zone, even if this code point is necessary to represent constructs typically used for identifiers. The root is a shared resource which is used by users from all types of backgrounds. Users of languages that do not use U+02BC as a letter cannot be expected to distinguish it from punctuation. It should be noted further, that the goal is to create Label Generation Rules that allow reasonable mnemonics, but that are not intended to cover the fullness of the writing systems the way ordinary plain text does. The decision whether to include a code point in the MSR or LGR thus is different from the decision of encoding a character in the first place.

The code point U+02BC is assigned to MODIFIER LETTER APOSTROPHE for which the Script property is “Common” suggesting that the character is used with multiple scripts. Nevertheless the use of this

modifier letter is essentially a borrowing of a convention from the Latin script. English and French (as well as a number of other languages) use the apostrophe for contractions and these can be common and mandatory for correct orthography. Nevertheless, U+0027 is excluded from the root. Several other languages use U+02BC with the Latin script, where it forms a letter of their alphabet. Yet, because of the concerns described, it will not be available in the root.

While it would be possible to restrict the use of this code point to the LGR repertoire tagged with the "und-Deva" language code, the Integration Panel must take into account that the Root Zone is a shared resource. Allowing the use of U+02BC in one context would create a precedent for other languages that use this code point. Such languages are predominantly using the Latin script. If the code point were allowed for the Latin script, it would not be possible in the root to restrict its use to any subset of languages.

For example, it would therefore become available to express English and French contractions that normally would be represented with U+0027 APOSTROPHE. In the opinion of the Integration Panel this would produce an unacceptable risk of confusion for the root.

In summary, the request for addition of U+02BC is rejected by the Integration Panel, because the code point is deemed unacceptable for the root due to being indistinguishable from basic punctuation.

Integration Panel Membership

This document is released by the Integration panel, composed of Asmus Freytag, Michel Suignard, Will Tan, Nicholas Ostler and Marc Blanchet.

References

[Procedure] Internet Corporation for Assigned Names and Numbers, "Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels." (Los Angeles, California: ICANN, March, 2013) http://www.icann.org/en/resources/idn/variant-tlds/draft-lgr-procedure-20mar13-en.pdf

[RFC5892] Faltstrom, P., Ed., "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA)", RFC 5892, August 2010.

[UAX24] UAX #24: *Unicode Script Property.* An integral part of The Unicode Standard. Most recent version available from http://www.unicode.org/reports/tr24/. Version 6.3 available as http://www.unicode.org/reports/tr24/tr24-21.html.

[Unicode63] The Unicode Consortium. The Unicode Standard, Version 6.3.0, defined by: "The Unicode Standard, Version 6.3.0", (Mountain View, CA: The Unicode Consortium, 2013. ISBN 978-1-936213-08-5). http://www.unicode.org/versions/Unicode6.3.0/.