IP Feedback on Latin LGR Proposal

DATE: 2019-02-02

The following is a somewhat consolidated compilation of separate reviews by individual members of the Integration Panel of the 2019-01-19 proposal draft by the Latin GP.

In some instances, our conclusions are tentative and may not represent the final consensus by the full Integration Panel. However, in the interest of time they are submitted here as a basis on which to conduct a planned joint discussion.

Also, for reasons of time, no attempt was made to comprehensively review all features or the presentation of the proposal in detail.

# Comments on Repertoire

While the IP is not reviewing the repertoire as such at this point, we do have a few comments and remarks:

1. Please document in Section 5.3 that à (A GRAVE) and ù (U GRAVE) are used in French; and also in any other place where language is associated with a specific code point.
2. There are some concerns with adding any code point with combining macron below (0061 + 0331, 0065 + 0331, 0069 + 0331, 006F + 0331, 0254 + 0331, etc…); at least with adding them without any variants. The problem arises from the widespread practice of underlining URLs, which hides these diacritical marks.
3. The same issue applies to code points with a combining dot below, which also can be entirely hidden by underlining (and which is reported to have been used on spoofing on the second level).
4. The GP might look into proposing a principle under which the effect of underlining should be considered when analyzing confusability. We saw that the point is mentioned in section 6.1 as ‘HTML underlining’ (not the best term for this) but we would like to see it more fully developed.
5. Please verify that the combining marks in the 3-codepoint sequence: 025B + 0331 + 0308 are in the correct order per NFC.

# Comments on Variants

## Variant set 1 - Alpha vs A.

NOTE: *all variant sets are as numbered as in the publication format HTML*

Marginal, but it may perhaps be worth discussing whether there exists a consistent rationale for this and similar cases.

There are some concerns about the proper way to analyze the case of ‘greek alpha’ versus ‘. It is true that alpha in many fonts looks rather close to the typical bowl-a glyph, which is the same as the glyph for 0251 LATIN SMALL LETTER ALPHA (ɑ) and which glyph is used in some fonts to represent regular ‘a’. (In particular, the same glyph is commonly used in cursive and italic style Latin fonts). Had 0251 been included in the Latin LGR it would have been possible to immediately analyze it as a variant of 0361 (Greek alpha). And at the same time, because in ordinary text practice users will accept both bowl and hook-a for ‘a’, it U+0251 would have been a candidate for a variant of ‘a’ – with transitive closure completing the set.

However, this direct analysis is unavailable, because U+0251 is not part of the required repertoire. Therefore, any analysis would have to establish a direct variant relation between ‘a’ and ‘alpha’.

There are certainly fonts (listed as intended for user interface use, such as Meyrio UI) which show Greek letterforms that any reader of Latin would likely accept as full substitutes (especially if the remainder of the FQDN is in some other script).

.ραρα vs .papa (screenshot: )

Presenting isolated code points side-by-side as was done in figure G01 in page 47 accentuates the distinction between their glyphs. It might therefore be less convincing because labels are typically produced with a single consistent font, not in ransom note fashion, and users do not always have immediate comparison.

In looking at the example above (in the font given), the Greek alpha definitely looks like a fully acceptable letter ‘a’ (just not the hook-a shape). However, in many other font one would say that the Greek alpha is just visually similar to a Latin bowl-a, and not perhaps not identical enough to justify a blocked variant.

Here the same labels are shown in standard italic font for this document: *.ραρα vs .papa*What can be seen is that the Latin label now uses a “bowl-a”, underscoring the fact that users will normally accept both hook and bowl shapes without question.

## Variant Set 10 - The ‘Circle’ Glyphs

LGR-3 includes a number of circle glyphs as cross-script variants; once that version is published, the GP will need to extend the analysis to add the other circles (beyond European scripts).

## Variant Set 11 - P and Rho

This set represents only near homoglyphs, but it may be worth discussing whether users do accept the substitution. It has been reported that there are existing in-script registrations that use LETTER WYNN which is about as close to P in appearance as Rho to spoof apple.com.

While WYNN and RHO differ in which elements deviate from the ‘standard’ shape of ‘p’, it doesn’t make RHO less attractive target for spoofing (other than that, as cross-script, it’s already a bit more restricted).

.αρρ .app (Meiryo UI)
.ρορ .pop (Segoe UI)

While fonts that omit the top left serif on the ‘p’ are not typical for user interface fonts, the substitution may be readily acceptable whenever a Latin TLD is used with non-Latin second level.

## Variant set 13 - R vs Ghe

Also not a true homoglyph, yet worth discussing on the basis of effective substitutability; this showed up in a recent sample of actual spoofed registrations for domains in .com.

## Variant Set 15 - U and look-alikes

This set gives rise to some concerns regarding possible overreach, particular since it has so many members.

## Variant Set 19 Beta vs. Sharp S

Not a homoglyph, but the claim that users will accept a substituted beta based on handwritten forms seems worth a discussion. Definitely English users would (and they happily type beta when they cannot find a sharp S). Note that italic fonts give a descender to the sharp-s, again priming readers to allow the substitution:

*Italic sharp s: ß --- Greek:β*

It may be the case that a typographer would always be able to point to some differences, but perhaps the standard of normal observant user does not rise to that level of discrimination:

. *νοβ .*vs .voß (Calibri)
.*νοβ* .vs .voß (Meiryo UI)

Notice that there is at least one font in which the Greek letter does not have a descender and that the Meiryo font shows a sharp, not rounded connection of the two loops of the ß; this leaves the subtle presence/absence of a small gap as the only truly distinctive feature.

## Variant set 16 - hook O vs. Sigma

The concern about this particular set stems from the fact that other scripts have backed off from similar cases where a continuing stroke or hook was generally different.

## Variant Set 34 – Dotless I vs. Latin and Greek Iota

This is a rather large set, also involving Armenian.

## Variant Set 37 - Schwa versus Latin turned e

In Section 6.4, the in-script variants 01DD and 0259 are claimed to be imposed by transitivity to 04D9. This is both correct and false. While the deferred Cyrillic LGR would indeed impose this variant by transitivity, 01DD and 0259 are inherently identical without a need to invoke their relationship to 04D9. This should be documented in a suitable way.

## Missing variant sets

We were rather expecting to see dot below and line below as variants of base letters, given how those two diacritics in particular disappear when URLs are underlined.

More troubling is that the submitted XML and document differ in the set of proposed variants. This makes any mechanical analysis by the IP useless, or worse, misleading. We request that this be avoided in the future.

## Small letter ‘f’ and “hook f” U+0192

This is another variant set depending on an argument that ordinary users accept an “extraneous” descender on the letter ‘f’, because such forms are common in *italic fonts*, among others.

In this context, it might be noted that for other scripts, the IP has accepted arguments that presenting certain distinguishing features outside their expected context will cause a security risk of the kind amenable to mitigation by blocked variant. Therefore, this case is worth a discussion.

## Small letter ae versus small ligature oe

IP may disagree with that one. Even in fonts that use the U+0251 form of ‘a’, there is always a clear design distinction between the ‘a’ form and the ‘o’ form in the two ligatures. In the ligature a vertical bar will be visible between the ‘a’ and the ‘e’ if you use the alpha variant form. And all italics forms preserve the unique form of the ae (does not look much like a ligature of the italic a and italic e in that case). This may be a case of mere visual similarity.

##  Combining Diacritics

Some reviewers agree with confusability between combining marks as introduced by the Latin GP. They don’t touch the subject that many Vietnamese labels would be totally confusable by non-Vietnamese readers (for example 1EA5 ấ and 1EA9 ẩ , in general similar appearance of hook and acute when used with another above combining mark)

The Latin GP needs to document this issue, even if it is different from the other diacritics case, where it is possible to substitute an *unexpected* diacritic usually used in a different language so that it is accepted by users as a native character to their language. The Vietnamese case would be establishing confusion within the repertoire of a single language – the expectation would be that readers are trained to watch for *expected* differences.

# Other comments

## Issues derived from scan of the XML

(1) the XML had one issue that forced us to edit it before our tool would accept it: on line 348, the ref 151 occurs twice

(2) the XML used type="out-of-repertoire" and not type="out-of-repertoire-var".

(3) reference 167 (Pulaar sound inventory) is unused; reference 149 (Fula, with alternate Spelling pulaar) is listed for U+01B4, but the language mentioned for ref 149 is not cited for U+01B4. These and similar issues may affect the table in section 5.3 of the main document as well and should be reviewed.

(4) variants currently use auto-generated comments not appropriate for the RZ-LGR, while others carry no comments. For consistency with other LGRs the following comment conventions are suggested:

* for cross-script homoglyphs: both mappings get the same comment: "cross-script homoglyph"
* for in-script variants: each mapping gets the appropriate comment, which may depend on the reason for the variant and contain the script identifier to mark it as in-script, for example: "Latin homoglyph" for U+01DD and U+0259
* In scripts where there's a mapping between nominal and final forms, the in-script comments would be something like "nominal form variant" and "final form variant", in each case describing the nature of the target of the mapping

## Miscellaneous Comments on the Document (editorial)

Repertoire, item 83

This seems to be Azeri, a Turkic language like the others cited, not Azer (which is a dialect of Soninke). [159] cited is an authority only on Azeri, not Azer.

p. 36 §6 para 2

In accordance to the Procedure ==> In accordance with the Procedure

p. 37 §6.1 para 2

underneath ==> under

p. 37, §6.1.1

clear cut ==> clear-cut

p. 38

Alphabet ==> alphabet

p. 39 §6.1.3.1 para 2

Many of such ==> Many such

p. 41 §6.2 para 3

wide-spread ==> widespread

p. 49 para second-from-last, end of 4th line

which is the is a ==> which is the

p. 50 , para 4

some historically established, ==> some historically established logos.

p. 52, para 2

lexem ==> lexeme

page 54, line 1

decisionmaer ==> decision-maker

p. 56, middle of page

Diaresis ==> Diaeresis

p. 59, §8

The list of contributors, possibly with relevant expertise, is still missing.