Dear IP,

The Latin GP appreciates the additional input received on August 29, 2018 titled “Diacritics below a security risk?”.

In its feedback, IP makes a case for potential security risks when certain diacritics below (e.g. dot and line below) are used in a domain name. The risk is not always apparent, but it reveals itself when the diacritic below is obscured by an underline, which is the typical formatting feature of hyperlinks. The IP asserts “**Of all diacritics, diacritics below can be difficult to distinguish or be prone to clipping”.**

Security risks deserve a place in our analysis, so Latin GP is committed to explicitly discuss this matter, resolve whether such risks constitute an actual security risk to the Root Zone and decide a solution vis-à-vis the LGR.

The GP is however unclear how many risks should be dealt with at the level of LGR vs. post-LGR review processes, such as String Similarity Review, DNS Stability Review processes and is therefore requesting additional guidance from IP regarding the recent communiquée which seem to be at odds with the earlier ones.

Specifically, the Latin GP notes these statements from the August 29 email:

*“It can be argued users have no working understanding of typography and would not reliably interpret small gaps or bulges in the underline as being related to an unfamiliar code point”*.

*“The IP would like to encourage the Latin GP […] to explicitly examine [the diacritics below] example and other cases like it, where code points can become indistinguishable in common usage scenarios for IDNs”*.

Thus far, the Latin GP has been aiming to deal with the “straight forward, non-subjective cases” of visual similarity by using a four-point visual similarity scale system. Per the LGR Procedure and additional IP feedback, the panel decided to admit a variant candidate when the source and target glyphs are visually identical; in the context of cross-scripts variants, the panel expanded the threshold to include nearly identical cases when the visual similarity can be attributed to differences in font design to account for indistinguishable differences in font design. Our methodology is explained in detail in our revised proposal.

In this context, the Latin GP wants to confirm that prior guidance from IP (below), which we find consistent with the LGR Procedure, is not at odds with the August 29 email.

                LGR Procedure:

“Finally, in investigating the possible variant relations, Generation Panels should ignore cases where the relation is based exclusively on aspects of visual similarity.”

October 18, 2017: Principles for repertoire and variants – Feedback from IP

*“In the context of the Root Zone, the Procedure is quite clear in that it considers simple similarity of appearance to be outside the scope of the Root Zone LGR.”*... *“Having the Root Zone exhibit fundamentally different design decisions with respect to variants than those found on the second level would have to be justified by strong arguments based on factors special to the Root Zone.”*

March 22, 2017: Latin GP Proposal: IP Feedback

*“The kinds of variants to be defined in the Root Zone LGR are limited to homoglyphs, which are characters essentially identical appearance by design, instead of merely similar appearance.”*

In light of the situation described above, the Latin GP kindly asks the IP to provide further guidance by way of the following cases:

1)

**0131 (ı) and (ɩ) 0269;**

**00E1 (ă) and (ǎ) 01CE;**

**006A (j) and (į) 012F;**

**00FC (ü) and (ű) 0171;**

**006B (k) and (ƙ) 0199;**

**0070 (p) and (þ) 00FE;**

**0064 (d) and (ɗ) 0257;**

**00F5 (õ) and (ō) 014D;**

**1E35 (ḵ) and (k) 006B;**

**0137 (ķ) and (k) 006B;**

**014D (ō) and (o) 006F;**

**1E37 (ḷ) and (l) 006C;**

**010B (ċ) and (c) 0063;**

**014B (ŋ) and (η) 03B7**

As of yet, the Latin GP has not found any(group of) language that uses these pairs of code points as the same (i.e. no evidence of interchangeability), which is why semantically, these sequences cannot be considered variants. Visually, however, the sameness or interchangeability of code points depends entirely on the user and while some users may be able to tell the difference, others may overlook the differences for psycho-linguistic or cognitive reasons, despite the fact that the glyph representing these code points clearly feature distinct elements of design as also demonstrated in non-representative test run by GP members with a couple dozen users, where users were even unable to spot differences when given clues to the kind of differences or specific contexts.

While a number of previous references and guidances seem to have suggested that the variant analysis on the visual side should be based on visual identity, the August 29 email, seems to suggest that the GP should consider at least cognitive factors and that by consequence, code points can be deemed variants on cognitive grounds even where they do not seem to fulfill simple criteria of visual identity because of clearly distinctive elements of design, such as diacritics.

Therefore the GP would like to request a fuller explanation on the view of the IP regarding potential or expected variant relationships in code points separated by other elements of visual design apart from modifiers below or at the baseline balance highlighted in the most recent communiqué by IP, as well as in how far the GP is allowed to deviate from such principles based on its own analysis.

2)

**006D (m) and 0072 006E (rn);**

**0062 006C (bl) and 004B (ы)**

In each of these two cases there is a sequence of letters that can be argued to be visually confusable with the target (single) code point in the set. Is it possible for sequences of characters or strings to be in a variant relationship with single codepoints, even where such sequences are not unique elements of the repertoire but can only occur as sequences of two independent elements in a label?

3)

**006C (m) and (м) 043C;**

**0074 0074 (tt) and 03C0 (π);**

**0074 (t) and 03C4 (τ),**

**0069 006F (io) and 044E (ю)**

In these cases, (sequences of) letters from the Latin script repertoire are visually distinctive from a letter from the Cyrillic and Greek script repertoire, respectively. However, the upper case form of the same Latin script letter would be clearly visually similar enough to be confusable by most users of the script-using community. The GP understands the upper case code points are not part of MSR-3 and therefore out of scope as elements of the repertoire itself. However, it is not immediately clear whether this rules out such considerations when it comes to the variant analysis, particularly given very concrete stability risks as probably all known user interface implementations of IDNA will consider an upper-case label equivalent with a lower case label, despite the fact that the lower case form is the only string an applicant can register. The GP would like to request the IP again to clarify on whether generally variant relationships by proxy (as in the case of upper-script forms to code points which are part of the repertoire) as well as specifically upper case in the context are out of scope and cannot be considered in the case of variant relationships and WLEs.

4)

(2) and 0577 (շ)

Here, a codepoint from the Armenian repertoire may be visually similar enough to be confusable with the numeral two, used also by the Latin-script using community. While numbers are not part of MSR-3 and therefore out of scope as well as excluded from the use in TLDs, second and further level zones may admit numbers for use in IDNA labels. Where subdomains occur or in other longer strings consisting of several label strings, a user may not be able to tell which string is the TLD identifier, whereby hypothetically a security risk may arise. The GP has discussed the present case and decided that while the numeral two is clearly out of scope, this may be a candidate for a variant rule at the point of integration of the Armenian LGR proposal, as those numerals are used by nearly all script using communities.

5)

**0069 (i) and (ı) 0131;**

**006C (l) and (լ) 056C;**

**006C (l)and (ι)03B9;**

**0077 (w)and (ω) 03C9;**

**006C (l) and (լ) 056C;**

**0062 (b) and (ъ) 044A**

In these cases, the pairs of code points show rather distinctive design features. However, some of these may be identified as hand-written pendants to the typed versions (and vice versa), with the typed versions having usually formed the basis for digital typography. Since the recognition of characters is strongly dependent not only on cognitive but also on psycho-linguistic factors inherent to users, this may lead to confusion, particularly since nearly all script users first start learning hand-written forms of any script before proceeding learn to both canonical shapes as well as variations in typed and digital typography. Therefore, despite apparent differing visual rendering in most digital fonts users may unconsciously mix-up the sets, as they have learnt to consider such forms equivalent, which may introduce security and stability issues to the zone. The GP would like clarification from IP on whether the current process permits representation of such variants by proxy due to psycho-linguistic and cognitive factors in variant rules and if IP has any expectation on where to draw the line.

The Latin GP looks forward to IP’s feedback.