Further IP Comments on Latin Variants

DATE: 2018-09-25

# Overview

In response to a discussion document from the Latin GP asking for clarification, the IP presents the following observations. It is hoped that the Latin GP will find these helpful in coming to a determination which variants to define and what rational to base that decision on.

# Nature of Latin Script

The IP agrees with the GP that the Latin script contains no or almost no *semantic* variants. Semantic variants imply a change in spelling without change of meaning. There are some languages, like English, where some spelling alternations (like the use of diacritics in words like “naïve”) or the use of ligatures (like “oe” and “ae”) may be optional. However, these cases do not appear to rise to a level that would be addressed as semantic variants the Root Zone LGR.

The hallmark of the Latin script is that it is used by so many different languages, each extending the basic set of 26 letters with additional ones, typically derived from the addition of a diacritic to an existing shape. (There are exceptions to both principles: not all languages use all 26 basic letters and some add additional basic shapes).

Generally, then, one can expect a native user of the Latin script to be aware that diacritics are an important means of distinguishing characters and to be able to detect the substitution of an unfamiliar one. This counteracts the issue that diacritics are generally small.

Not all users are familiar with all diacritics or extensions to character shapes like hooks and tails. Generally, one would expect users to only be familiar with the set used in their language. Users who are only familiar with English or with the basic Latin alphabet, be it because they are native English speakers or use Latin letters only as modern adjuncts to their native scripts would be expected to be unfamiliar with both diacritics and shape extensions. No users can be expected to distinguish IDNs in an entirely unfamiliar script. That would imply that the first concern should be with users interacting with TLDs that look like they are in their native language and writing system.

For some languages, there are obvious cognate words that are spelled the same, except for differences in diacritics or use of some special letter. Scandinavian languages share many words where only the vowels are spelled differently (using a-diaeresis instead of ae ligature, or o-diaeresis instead of o-slash). Likewise, the different nationalities using German will either use or not use the ß.

The VIP project’s study group did come out against the idea of defining variants for the Latin script; however it should be noted that at the time that work was done, the focus may have been more on the question of allocatable variants, which are usually semantic variants that should both be available to the same requester.

Words in the Latin script can be quite long; the only IDN TLDs delegated in the Latin script so far are “Vermögensberater“ and “Vermögensberatung“. If a significant percentage of Latin letters were to have variants, then the number of variant labels for such cases would grow quickly, with no viable scheme for reducing such combinatorial explosion. The IP therefore would tend to consider allocatable variants inappropriate and infeasible for the Latin script, which effectively agrees with the position by the VIP Study Group that there is no scope for allocatable variants in the Latin Scripts.

# The Nature of Variants

While allocatable variants have fundamental issues due to the fact that they seemingly change the exact match nature of the DNS, and therefore need to be tightly limited, no such concerns apply to blocked variants.

In fact, where it comes to blocked variants, there has been a bit of refinement in thinking. Blocked variants favor the first-delegated label over any of its variants. They can be a powerful tool to reduce the security risk from mischievous applications for look-alike labels.

For the purpose of detecting collision between an applied for label and all the variants of a delegated label, there exist very effective optimization strategies that make the number of blocked variants per label largely irrelevant. Furthermore, studies done by the IP using sample Root Zone LGRs and large corpora of putative labels have shown that for the languages and scripts studied, even extensive definitions of blocked variants did not result in noticeable reduction of the overall namespace. These tests included the Ethiopic script where the IP was able to study the effect of variants implemented for one language on other languages using the script.

Therefore, generally, the IP would find the use of blocked variants a viable tool to address certain security issues; however, limitations apply.

The Root Zone LGRs are supposed to be applied mechanically, and their outcome not subject to appeal. That implies that cases that are not “cut & dried” or otherwise manifestly obvious are a poor fit for being addressed in the design of the LGR. That goes for blocked variants as well.

Formally, as used in the Root Zone, a variant relation is an equivalence relation. It has to be symmetric and transitive. For the optimizations mentioned above, this symmetry and transitivity has to hold not only for the code point variants as defined, but also for the computed variant labels: all variant labels for a given label must also be variants of each other (and of no other label). (This is can become an issue if variants are defined for some types of sequences).

Further to the nature of variants as equivalents: this also means that the variant relation is best used for cases where each of the variants are fully substitutable to the user.

In the case of the Latin script, the most obvious type of fully substitutable pairings is that of homoglyphs, for example ǝ and ə. No user can tell these two apart, and if two labels only differ by the choice of code point (U+01DD vs U+0259) they are effectively indistinguishable and must not be allowed to be both delegated. Those two code points therefore make an obvious target for defining a blocked variant.

Their relation is that of “exact homoglyph”; such cases clearly should not be subject to appeal and therefore the use of blocked variants in the LGR would be appropriate on that reason as well.

In contrast, cases where letters are, in fact, distinguishable, there enters some subjectivity in the analysis – the more, the less the degree of actual similarity in a given case. The more subjective a decision becomes, the less it is appropriate to enshrine it in the LGR, and the more it naturally becomes a matter for appeal. That is the reason why visual similarity is normally out of scope for the RZ LGR.

In addition, merely similar, but distinguishable labels a formally a bad fit for the variant mechanism. That is, because with only moderate similarity, a code point may have an “intermediate” shape where it is similar to two different code points, but the two, being further apart in perception space, may well be fully distinguishable.

Where the relation between code points changes from visual equivalence to similarity, the variant mechanism stops being applicable, because the latter can only model equivalence relations. (See also RFC 8228).

# Security, Mitigation and Variants

There are some instances of “near homoglyphs” and other cases where code points may be homoglyphs under certain circumstances, that pose strong security risks and where it is a matter of principled decision as to whether the treatment as variants is possible, and if so, whether the benefits outweigh the drawbacks.

This decision is up to the Latin GP. Under the process set up by the Procedure the GPs have the expertise when it comes to the nature of their script and are therefore tasked to propose any variants and to provide the rationale for their decision. The IP is tasked with review (which will take into account the facts presented, as well as the rationale, but also will consider issues such as consistency across the Root Zone and whether the proposed LGR promises to keep the DNS secure).

Generally, code points that differ only by the presence/absence of a diacritic above would seem easily distinguishable (even though English users may treat letters like i-diaeresis as optional in their spelling, e.g. of “naïve”).

However, the IP had asked the question whether the common practice of underlining would defeat the use of certain diacritics below in distinguishing labels (most affected appear to be the dot-below and macron-below). If the GP agrees that these labels become *indistinguishable* in a very common scenario related to use of IDN TLDs, then it may decide that in that instance the code points are substitutable without the user being able to tell, and that therefore they do behave like variants.

Going further, there are some pairs of diacritics that, while nominally different, are harder to tell apart and commonly not used in the same language (like a-breve and a-caron, or comma-below and cedilla-below). In some cases, not all fonts will show the nominal shape, for example, when this issue was studied by Unicode, many fonts were found that implemented cedilla below using a comma-below shape. For typographical reasons, diacritics below may be less distinctly rendered, a factor that the Latin GP may consider in its analysis.

Fonts can be very variable in the way they implement some typographic features. Generally, it is not enough to show that two code points are distinct or not in some fonts. However, when code points are consistently indistinguishable or nearly indistinguishable to an observant user in fonts that are commonly used for user interfaces that fact may be of relevance in making decisions on variants.

In these and similar cases, it would be up to the Latin GP to evaluate such near-homoglyphs and to decide whether the relation is so unambiguous and the risk so severe that addressing it in the LGR is both feasible and appropriate. Such decision would be provided with a rationale adducing the evidence on which the GP made its decision, so that the IP can review it as discussed above.

In the case of the Devanagari script, the GP argued convincingly that the practice of applying a nukta (dot below) to certain vowel characters as used in some minority language would be unexpected to the majority of users of the script and would lead to making the vowels with and without nukta effectively indistinguishable. (In this context, it may be worth noting that the nukta typically is even smaller in relation to the base character than a dot-below diacritic in Latin). The GP opted to define vowels with and without nukta as blocked variants of each other.

Note that this argument is not the same as claiming that not all users can distinguish every character.

# Out of scope

Some example can be considered out of scope for one reason or another.

Generally, the IP is not considering UPPERCASE presentation of IDNs (which are inherently lowercase).

More importantly, because of the way upper and lower case shapes are related across scripts, it is not feasible to create variant relations. For example, symmetry and transitivity would force a variant relation between Latin “n” and “v” if the uppercase form of Greek “NU” was considered.

Speaking of variants between basic Latin letters: these are completely out of scope because they do not apply to non-IDN TLDs. For cross-script variants, any proposed variants that (via symmetry and transitivity) would impose an in-script variant relation in any non-Latin script would face a very high bar.

For the Root Zone, any relations that involve code points not in MSR-3 are also best not considered – unless it is intended to apply for these code points to be added for MSR-4.