Proposal for a Latin Script Root Zone LGR

Minority Report

Bill Jouris

The Latin Generation Panel has done an enormous amount of work, in an extremely complex area. Unfortunately, the end product is, in my opinion, seriously flawed.

Repertoire

The whole point of the IDN project is to make sure that users worldwide can create domain names in their own language and their own script. That includes not just very widespread scripts, but also very local ones. Of the 450+ living languages which use the Latin script, the Latin GP has chosen to include less than half in establishing the repertoire. (The omission is made worse by the fact that the vast majority of the languages which are not included are from sub-Saharan Africa, while none of the Panel members are from that region nor speak the languages used there.)

The criteria used were to include “official languages” plus languages which are not so designated but have at least 1 million users. However, classification as an official language depends on local politics, the resources available to the local government to conduct business in additional languages, and vagaries of colonial boundaries drawn by diplomats thousands of miles away without references to the peoples on the ground. And the threshold is both arbitrary and subject to the ever-changing populations of native speakers. Due to the criteria used, we have the ludicrous situation where a language like Hawaiian (25,000 native speakers) is included, while a language like Obolo (300,000+ native speakers) is not.

The Latin GP needs to go back and spend the few weeks required to include those languages. (I recognize that this will also require some additional work to evaluate possible variants involving the additional codepoints from the additional languages. So be it.)

Variants

One of the foundational principles for the Internationalization of Domain Names project is the

> **Least Astonishment Principle**: A Code Point in the Zone Repertoire should not present recognition difficulties to the zone's intended user population and should not lend itself to malicious use.[1]

Further, the IDN Variant TLD Implementation[2] document states that

[1] https://www.icann.org/en/system/files/files/lgr-procedure-20mar13-en.pdf

[2] https://www.icann.org/resources/pages/idn-variant-tld-implementation-2018-07-26-en

> The variant management mechanism should “promote a good user experience”, which means that it should “avoid including variant TLDs in a manner that would create user vulnerabilities or a probability of confusion”.

But the Panel seems to have lost track of that in establishing variants.

The Latin script makes use of some 20 diacritic marks, which are used to modify one of more underlying letters. However most languages use only a small subset of these. While the members of the Panel are, of course, familiar with all of the diacritic marks, the typical user will only be acquainted with half of them or less. This increases the potential for confusion, as it is difficult to recognize as different a diacritic which one does not even realize exists. For example, suppose the user has never encountered the Dot Above diacritic, and is expecting an Acute (or Grave) diacritic. When presented with a Domain Name which uses a Dot Above, he is unlikely to realize that it is not the Doman Name featuring an Acute that he was expecting.

Evaluation of possible variants involved the Panel members looked at pairs of glyphs side by side (a luxury that an end user would not have during normal use), while knowing that they were looking at two different glyphs. They then decided how likely it was that, in looking at a Domain Name, they would realize that the two were different.

The Panel initially took a simple majority rule approach (4 of the 7 Panel members). But the Integration Panel quickly advised the Panel that identifying variants was not a matter for a vote. What the Panel chose to do in response was to raise the bar, and only recognize as variants those that 5 of the 7 expert members found indistinguishable. What the Panel should have done at that point was to recognize that typical users have far less expertise and experience than the Panel members. And that our experts were doing a side-by-side comparison (a luxury the user would not typically have) of two glyphs that we already knew were different. And so if 3, or even just 2, members found the glyphs confusable in those circumstances, the typical naïve user looking at a domain name would also.

The criteria used has resulted in the Panel having an official position that some pairs of glyphs, which a *majority* of the members could not distinguish, nonetheless are different enough that a “reasonably careful user” would somehow magically see the difference. It is simply not possible to reconcile this with the principles above.

The Panel needs to go back and apply a more reasonable standard for variants. Fortunately, the spreadsheets used for the existing ratings are still available. Thus the revision should be quite straightforward, if tedious.

Underlining

Also, there is the issue of Underlining. The Panel asserts that the underlining added to domain names is interrupted by blank pixels on either side of any below-the-line diacritics, and thus a “reasonably careful user” will be able to find them. Even assuming that said user could spot diacritics that he has never seen and so doesn’t know to look for, there is another problem. While some browsers (Chrome, Firefox) do put in blank pixels, others (Internet Explorer) do not. Further, some widely used word processors (Word, PowerPoint, Excel) do not. It may be noteworthy that the software that ICANN uses in the Public

Comment feature also does NOT insert blank pixels. Thus it seems mistaken to say that users will generally be able to see below-the-line diacritics. In fact, any codepoint involving a below-the-line diacritic which is not connected to the basic letter should be considered a variant of that letter.

Other

There are also a couple of areas where the Latin GP was handed a decision from above, but which should be noted:

Capitals

Because domain names are strictly lower case, the various Generation Panels were directed to ignore capital letters when identifying variants. But this ignores the reality that users have decades of experience which teaches them that, in a domain name, upper and lower case letters are completely interchangeable. (The Greek GP has written about this as well.) For example, .COM would be entirely interchangeable with .com. Therefore, if a user encounters a Cyrillic TLD of .сом, he will automatically assume he is seeing .com in upper case. What he will NOT do is recognize, as the IDN project apparently expects, that the third letter doesn't look like a lower case M, and therefore this is obviously different from .com.

One rationalization for ignoring capital letters anyway is that the size of the letters will indicate that they are not capitals. Note, however, that the Root Zone LGR includes this: "In typical user interface fonts, even code points like "s" and "ട" (U+0D1F) may look indistinguishable." So, we have a type case for cross-script variants where, in the same font size, on letter is TWICE the size of the other. So such argument cannot stand.

The Latin GP cannot change this policy. But the IDN project should be asked to rethink it.

Numerals

Numerals are allowed in Second Level Domain Names, but not in Top Level Domain Names. Because the GPs are focused entirely on TLDs, they necessarily ignored numerals. But there are some potential variants here. For example:

Numeral One vs Latin Small Letter L

Numeral Three vs Latin Small Letter Ezh

Numeral Five vs Malayalam Letter Tta

Numeral Zero vs Malayalam Letter Ttha

While ICANN has been working on policies for variants in SLDs, there is no sign that the IDN project contemplates a new Panel to look at numerals as variants. So the Latin GP should at least mention the issue.